La mente come struttura del mondo

Il radicamento causale, la geometria dell'esperienza e i limiti dell'intelligenza artificiale

Saggio filosofico in continuazione di
«Intelligenza, Rappresentazione e il Confine dell'Orizzonte Computazionale»

I. Ciò che il meccanismo non può contenere

L'intelligenza artificiale contemporanea ha raggiunto una soglia che la storia intellettuale dell'Occidente non aveva mai attraversato, e non è la soglia che si proclama nei comunicati stampa: non è la soglia della performance umana su questo o quel benchmark, non è la soglia del linguaggio generato che supera il test di Turing in condizioni controllate, non è la soglia dell'immagine sintetica indistinguibile dalla fotografia. È una soglia epistemica più profonda e meno celebrata, e consiste nel fatto che per la prima volta gli esperimenti artificiali costringono la filosofia della mente a fare i conti con la propria imprecisione — a scoprire che le sue distinzioni erano, in molti casi, meno nette di quanto si pensasse, e che le sue intuizioni erano, in altri casi, più robuste di quanto l'ottimismo tecnologico concedesse.

Il saggio precedente che questa ricognizione prolunga aveva percorso il terreno con insolita ampiezza, attraversando quasi ogni tradizione del pensiero occidentale che avesse qualcosa di rilevante da dire sul rapporto tra intelligenza, rappresentazione e computazione. Il saggio che qui si intraprende sceglie deliberatamente la direzione opposta: la concentrazione su un'unica questione, la più difficile e la più irrisolta, quella intorno alla quale tutte le architetture analizzate nell'opera precedente si dibattono senza risolverla e che le critiche filosofiche di Putnam, Kripke, Tarski e Gödel circoscrivono senza esaurire. Quella questione è la seguente: in che senso un sistema materiale — neurale, computazionale, o di qualsiasi altra natura — può avere un rapporto con il mondo che sia qualcosa di più di una correlazione causale cieca, che sia comprensione nel senso in cui quella parola è usata quando si dice che una persona comprende il peso di una promessa, la struttura di una dimostrazione, il significato del silenzio di qualcuno che si ama?

Non è una domanda retorica. Ha un corpo tecnico preciso, che questo saggio si propone di costruire con rigore matematico dove è possibile, e con onestà concettuale dove la matematica non arriva ancora. La risposta che si tenterà non è una risposta di chiusura — una dimostrazione che il programma dell'intelligenza artificiale sia destinato al fallimento, o al contrario che il suo successo sia garantito dall'andamento dei benchmark. È una risposta di strutturazione: un tentativo di mostrare dove si trova il problema, di che tipo di problema si tratta, e quali risorse concettuali e formali sono disponibili per avvicinarsi a comprenderlo senza la pretesa di risolverlo.

L'argomento si svolge in tre fasi, ciascuna corrispondente a una sezione del saggio. La prima fase, che questa sezione inizia a costruire, riguarda la struttura del radicamento causale: non il generico appello all'incarnazione che la letteratura evoca troppo spesso come talisman filosofico, ma una ricostruzione matematicamente precisa di cosa significhi che una rappresentazione sia radicata nel mondo in senso causale, con derivazioni che mostrano dove le architetture attuali si fermano e perché. Il secondo momento riguarda la geometria dell'esperienza: l'analisi di come gli organismi biologici costruiscono rappresentazioni del mondo fisico attraverso l'interazione motoria, e di come questa costruzione abbia proprietà — in primo luogo la proprietà di rispettare la

struttura di simmetria dello spazio fisico — che nessun sistema addestrato su osservazioni passive può garantire di acquisire. Il terzo ambito riguarda i limiti formali: una rivisitazione dei teoremi di incompletezza e di inesprimibilità non come argomenti di impossibilità assoluta ma come vincoli strutturali che definiscono l'architettura che qualsiasi sistema cognitivo deve avere se vuole valutare autenticamente le proprie rappresentazioni.

Il filo che attraversa i tre movimenti è questo: il problema dell'intelligenza artificiale non è un problema di potenza computazionale. Non è risolto aumentando i parametri, ampliando il corpus, raffinando l'architettura a trasformatore, aggiungendo modalità sensoriali. È un problema di struttura del rapporto tra sistema e mondo, e la struttura rilevante è di natura matematica — non la matematica dell'ottimizzazione su distribuzioni fissate, che l'apprendimento profondo padroneggia con maestria crescente, ma la matematica dei gruppi di simmetria, dei grafi causali e della logica dei sistemi autoreferenti.

La distinzione su cui tutto si regge — e che la monografia precedente accostava in molti luoghi senza renderla il fuoco esplicito dell'analisi — è quella tra covariazione e causalità, tra sistema che risponde a correlazioni e sistema che comprende strutture. Un sistema che risponde a correlazioni può raggiungere performance straordinarie sulle distribuzioni in cui le correlazioni rilevanti sono abbondantemente rappresentate; fallisce sistematicamente fuori da quelle distribuzioni, e fallisce in modi che rivelano la struttura di ciò che manca. Un sistema che comprende strutture causali generalizza in modo qualitativamente diverso: perché le strutture causali sono invarianti rispetto a un insieme di trasformazioni molto più ampio delle correlazioni statistiche, e il sistema che le padroneggia porta con sé, in qualsiasi contesto nuovo, un modello del mondo che predice correttamente le conseguenze degli interventi — non soltanto le probabilità delle osservazioni.

Questa distinzione non è nuova. Hume l'aveva intravista nel suo scetticismo sulla causalità — riconoscendo che l'impressione di causalità è un abito mentale proiettato sulla successione regolare, non una percezione diretta di una connessione necessaria nel mondo. Kant l'aveva riformulata in termini della struttura categoriale dell'esperienza — la causalità non come dato empirico ma come condizione di possibilità dell'esperienza organizzata. Reichenbach l'aveva trattata nell'epistemologia della fisica, connettendola alla struttura delle teorie scientifiche come sistemi di leggi controfattuali. Pearl l'ha infine resa computazionalmente precisa attraverso il do-calcolo, che separa formalmente la distribuzione osservazionale — ciò che si vede succedere quando si osserva il mondo senza intervenire — dalla distribuzione interventistica — ciò che succederebbe se si intervenisse sul mondo modificando attivamente una delle sue variabili.

La precisione del do-calcolo è il punto di partenza tecnico di questo saggio, e vale la pena richiamarla con più dettaglio di quanto la monografia precedente non avesse fatto, perché è nel dettaglio tecnico che la filosofia trova i propri vincoli più severi.

Sia dato un grafo causale aciclico diretto G = (V, E), dove V = $X_1, \ldots, X$ è l'insieme delle variabili e E è l'insieme degli archi orientati che rappresentano le relazioni causali dirette tra di esse. La distribuzione osservazionale congiunta $p(x_1, \ldots, x)$ è consistente con G se fattorizza secondo il criterio di Markov causale: ciascuna variabile è condizionalmente indipendente dai propri non-discendenti in G dato il proprio insieme di genitori, $Pa(X_i)$. Questa fattorizzazione può essere scritta come

$$p(x_1, \ldots, x) = \prod_{i=1}^{n} p(x_i \mid pa(x_i))$$

dove $pa(x_i)$ denota la realizzazione specifica dei genitori di X_i. Fin qui, nulla di diverso da un modello grafico probabilistico standard. La novità del do-calcolo è l'operatore di intervento $do(X = x)$, che rappresenta un'azione sul mondo che fissa il valore di X a x indipendentemente dai fattori che lo determinano ordinariamente. Sotto l'intervento $do(X = x)$, il grafo viene mutilato: tutti gli archi entranti in X vengono recisi, e la distribuzione interventistica è

$$p(x_1, \ldots, x \mid do(X = x)) = \prod_{i \neq j} p(x_i \mid pa_{\mathcal{G}_X}(x_i)) \cdot \mathbf{1}[x = x]$$

dove $\mathcal{G}_X$ è il grafo privato di tutti gli archi entranti in X .

La struttura formale di questa operazione rivela immediatamente perché la distribuzione interventistica non sia recuperabile dalla distribuzione osservazionale senza conoscenza del grafo causale. Due grafi causali diversi possono essere Markov-equivalenti — generare la stessa distribuzione osservazionale — pur essendo causalmente diversi nel senso che producono distribuzioni interventistiche radicalmente diverse. Un sistema che apprende la distribuzione osservazionale senza modellare la struttura causale che la genera non può distinguere tra grafi Markov-equivalenti, e quindi non può, in generale, calcolare distribuzioni interventistiche corrette. Questo è il teorema di non-identificabilità causale nel suo nucleo più essenziale: l'informazione causale non è contenuta nella distribuzione osservazionale, e non può esservi estratta senza assunzioni aggiuntive sulla struttura del sistema che la genera.

Le assunzioni aggiuntive che permettono l'identificabilità — ciò che Pearl chiama le condizioni di do-identification — sono tre: l'assenza di confondenti non osservati tra certe coppie di variabili (condizione di sufficienza causale), la disponibilità di variabili strumentali appropriate, o la presenza di variazioni di dominio che modificano certi meccanismi causali mantenendone altri fissi. Quest'ultima condizione — la variazione di dominio — è quella che connette il framework di Pearl con i risultati di identificabilità degli iVAE di Khemakhem e collaboratori, e merita una trattazione separata nel secondo movimento di questo saggio. Qui importa sottolineare un punto che la monografia precedente non aveva reso abbastanza esplicito: le condizioni di identificabilità causale non sono condizioni sulle prestazioni del sistema — non si è i condizioni sulla quantità di dati, sulla potenza del modello, sulla qualità dell'ottimizzazione. Sono condizioni sulla struttura del problema, e se la struttura del problema non le soddisfa, nessun miglioramento del sistema risolve il problema. Questo è il senso preciso in cui la causalità pone vincoli strutturali, non contingenti, all'apprendimento.

Il non detto che percorre gran parte della letteratura sull'intelligenza artificiale — e che questo saggio vuole rendere esplicito con la chiarezza che merita — è che i sistemi di apprendimento profondo operano quasi interamente al primo livello della ladder of causation di Pearl: il livello dell'associazione. Questo livello risponde a domande della forma «cosa osservo?» attraverso l'apprendimento di distribuzioni condizionali sulle variabili osservabili. Il secondo livello — l'intervento — risponde a domande della forma «cosa succederà se agisco?» attraverso il calcolo di distribuzioni interventistiche. Il terzo livello — il controffattuale — risponde a domande della forma «cosa sarebbe successo se le cose fossero andate diversamente?» attraverso il ragionamento su mondi possibili. I sistemi attuali, compresi i più sofisticati tra i World Models analizzati nella monografia precedente, operano al primo livello con maestria crescente; toccano il secondo livello in modo parziale e incerto nei sistemi di reinforcement learning che eseguono azioni reali nel mondo; non affrontano sistematicamente il terzo livello.

Questa gerarchia non è un giudizio normativo: non si afferma che i sistemi di primo livello siano inferiori o inutili. Sistemi che operano al primo livello con alta qualità sono straordinariamente utili, e la loro utilità cresce sistematicamente con la scala. Si afferma invece qualcosa di più preciso: che la comprensione nel senso pieno che distingue l'intelligenza umana — la capacità di ragionare su ciò che succederebbe se il mondo fosse diverso, di pianificare attraverso catene di interventi causali, di attribuire responsabilità e di rivedere le proprie credenze alla luce di nuova evidenza causale — richiede operazioni di secondo e terzo livello che i sistemi puramente associativi non possono garantire di aver acquisito, indipendentemente dalla scala.

La tesi che l'aumento di scala non produca comprensione causale incontra la sua sfida empirica più diretta nel fenomeno delle emergent capabilities nei modelli linguistici su larga scala — un fenomeno che la monografia non tematizza esplicitamente ma che mette sotto pressione l'argomento in un modo che richiede risposta precisa. Il fenomeno è stato documentato sistematicamente da Wei e collaboratori nell'ambito del BIG-Bench benchmark (2022) e successivamente da Ganguli e collaboratori: esistono capacità cognitive — ragionamento aritmetico multi-step, decomposizione analogica, comprensione di strutture sintattiche complesse, ragionamento su controfattuali — che compaiono discontinuamente, in modo apparentemente non previsto dall'estrapolazione delle leggi di scala calcolate per capacità più semplici, quando la dimensione del modello supera soglie parametriche specifiche. Queste capacità non erano presenti, neppure in forma degradata, nei modelli più piccoli; compaiono bruscamente, con una pendenza nella curva di apprendimento in funzione della scala che diverge da quella delle capacità più semplici.

La questione che questo fenomeno pone all'argomento della monografia è la seguente: se la comprensione causale è una struttura di tipo qualitativamente diverso dal pattern-matching statistico — se è un funtore globale piuttosto che un'approssimazione locale — perché non è lecito interpretare l'emergenza discontinua di queste capacità come la manifestazione di una transizione di fase analoga, in cui il sistema supera la soglia al di sopra della quale le strutture causali locali si coordinano in strutture causali globali? La risposta a questa domanda non può essere semplicemente assertiva — non basta affermare che le capacità emergenti sono ancora pattern-matching statistico sofisticato — ma richiede un argomento che distingua tra performance emergente e comprensione genuina in modo che non sia circolare.

L'argomento disponibile ha due mosse. La prima è concettuale: le capacità emergenti identificate nei benchmark — ragionamento aritmetico, comprensione sintattica, analogie — sono valutate attraverso metriche di performance su distribuzioni di task, non attraverso test di generalizzazione interventistica nel senso del do-calcolo. Un sistema che eccelle su un benchmark di ragionamento aritmetico ha imparato a produrre le risposte corrette per le istanze del benchmark; non ha necessariamente acquisito un modello causale dell'aritmetica che gli permetta di rispondere correttamente a interventi arbitrari sulla struttura del problema. La distinzione tra performance su distribuzione e generalizzazione interventistica è precisamente la distinzione tra il primo e il secondo livello della ladder of causation di Pearl, e il fatto che le capacità emergenti vengano valutate esclusivamente al primo livello rende la loro rilevanza per la tesi causale non determinata dall'evidenza disponibile.

La seconda mossa è empirica: studi di robustezza sistematica, come quelli condotti da Dziri e collaboratori sulle capacità di ragionamento compositivo dei transformer (2023), mostrano che le capacità emergenti sono fragili rispetto a variazioni superficiali della formulazione dei task che non modificano la struttura causale del problema. Un sistema che risolve correttamente una variante di un problema di ragionamento può fallire sistematicamente su una variante formulata diversamente ma strutturalmente equivalente — il che suggerisce che il sistema ha appreso features superficiali correlate con la soluzione corretta nella distribuzione di addestramento, piuttosto che la struttura causale profonda che permetterebbe la generalizzazione a formulazioni nuove. Questo pattern di fragilità — la cui firma

sperimentale è la degradazione della performance che è funzione della distanza superficiale dalla distribuzione di addestramento piuttosto che della distanza strutturale causale — è la manifestazione comportamentale della distinzione tra approssimazione locale e funtore globale che la Tesi IV articola formalmente.

Rimane tuttavia aperto un problema metodologico che occorre riconoscere onestamente: la distinzione tra «performance emergente» e «comprensione genuina» rischia di essere circolare se «comprensione genuina» viene definita come ciò che i sistemi attuali non hanno. Per evitare questa circolarità, la distinzione deve essere ancorata a una caratterizzazione positiva e indipendente della comprensione causale — caratterizzazione che la struttura funtoriale proposta nella sezione V fornisce in termini formali: un sistema implementa comprensione causale genuina nella misura in cui il diagramma tra situazioni e rappresentazioni commuta, cioè nella misura in cui le trasformazioni causali del mondo si riflettono fedelmente in trasformazioni struttura-preservanti delle rappresentazioni del sistema. La domanda empirica è se i sistemi con capacità emergenti mostrano questa commutatività in modo sistematico e robusto fuori distribuzione, e l'evidenza disponibile suggerisce che non la mostrano — ma questa è una questione sperimentale aperta, non una conclusione garantita dalla struttura dell'argomento.

La domanda che rimane aperta, e che questo saggio si propone di articolare con la precisione che merita, è questa: quali architetture, quali regimi di addestramento, quali strutture rappresentative permetterebbero ai sistemi artificiali di acquisire capacità autentiche di secondo e terzo livello? Non si darà una risposta definitiva a questa domanda — sarebbe disonesto farlo dato lo stato attuale della ricerca. Si mostrerà invece che la risposta richiede innovazioni di tipo concettuale prima che ingegneristico, e che quelle innovazioni concettuali sono già in parte disponibili nella letteratura, sebbene non ancora integrate in modo organico.

II. La geometria dell'interazione: dal dato sensoriale alla struttura causale incorporata

Il problema del radicamento causale, come si è appena argomentato, non è un problema di quantità di dati: è un problema di tipo di interazione con il mondo. I sistemi biologici — e in primo luogo i mammiferi, ma con gradazioni significative anche gli invertebrati più complessi — non apprendono il mondo osservando le distribuzioni delle proprie osservazioni sensoriali passive. Lo apprendono attraverso l'azione e le sue conseguenze, attraverso un ciclo percettivo-motorio in cui il sistema emette un'azione nel mondo, osserva la risposta del mondo a quell'azione, e aggiorna le proprie rappresentazioni in modo che le future predizioni delle conseguenze delle azioni siano

più accurate. Questo ciclo — che nelle sue versioni più sofisticate include la pianificazione, l'anticipazione, la generazione di modelli controfattuali del tipo «se avessi fatto X invece di Y, cosa sarebbe successo?» — è la struttura fondamentale dell'apprendimento incarnato, e la sua differenza dal semplice apprendimento statistico passivo non è di grado ma di tipo.

L'incarnazione non è, in questa prospettiva, una metafora romantica per la «concretezza dell'esperienza» contrapposta all'«astrazione del simbolo». È una struttura tecnica precisa che ha conseguenze matematiche verificabili. Il punto di partenza è la seguente osservazione: un organismo che agisce nel mondo fisico tridimensionale apprende, attraverso l'interazione ripetuta con quel mondo, la struttura delle trasformazioni che il mondo realizza sulle sue osservazioni in risposta alle sue azioni. Queste trasformazioni hanno una struttura di gruppo — nel senso algebrico preciso del termine — che è determinata dalla geometria dello spazio fisico, non dalle contingenze statistiche dei dati di addestramento. La struttura di gruppo non è qualcosa che il sistema impone dall'esterno al mondo; è qualcosa che il mondo impone dall'interno al sistema attraverso la regolarità delle proprie risposte alle azioni.

Sia $\mathcal{O}$ lo spazio delle osservazioni sensoriali disponibili a un agente incarnato, e sia $\mathcal{A}$ lo spazio delle azioni che l'agente può eseguire. La dinamica dell'interazione tra agente e ambiente definisce una mappa — chiamiamola funzione di transizione — $T: \mathcal{O} \times \mathcal{A} \to \mathcal{O}$, che associa a ogni coppia (osservazione corrente, azione eseguita) la nuova osservazione risultante. In un ambiente fisico reale, questa mappa non è arbitraria: è vincolata dalla geometria dello spazio fisico nel quale l'agente si muove. In particolare, le trasformazioni dell'ambiente corrispondenti ad azioni nello spazio fisico tridimensionale formano una struttura di gruppo che è omotopicamente equivalente al gruppo euclidiano $E(3) = \mathbb{R}^3 \rtimes O(3)$ — il gruppo dei moti rigidi dello spazio fisico.

Questo punto merita di essere sviluppato con rigore, perché è il nucleo matematico del problema dell'incarnazione. Il gruppo euclidiano $E(3)$ è il gruppo semidiretto di $\mathbb{R}^3$ (traslazioni) con $O(3)$ (rotazioni e riflessioni). Il sottogruppo delle isometrie orientazione-preservanti — i moti rigidi propri — è $SE(3) = \mathbb{R}^3 \rtimes SO(3)$, dove $SO(3)$ è il gruppo speciale ortogonale delle rotazioni in tre dimensioni. Un elemento $g \in SE(3)$ può essere rappresentato come una coppia (t, R) con $t \in \mathbb{R}^3$ e $R \in SO(3)$, con la legge di composizione

$$(t_1, R_1) \cdot (t_2, R_2) = (t_1 + R_1\, t_2,\ R_1\, R_2)$$

che riflette la non-commutatività del gruppo: la composizione di due moti rigidi dipende dall'ordine in cui vengono eseguiti, perché la parte rotazionale di ciascun moto agisce sulla parte traslazionale dell'altro.

L'importanza di questa struttura per il problema dell'incarnazione emerge attraverso il seguente argomento. Supponiamo che un agente osservi una sequenza di osservazioni $o_1, o_2, \ldots, o_T$ prodotte dall'interazione con un ambiente fisico, dove l'agente esegue azioni $a_1, \ldots, a_{T-1}$ che corrispondono a moti fisici nello spazio. Ciascuna osservazione o_t è una funzione dell'orientamento e posizione dell'agente al tempo t, che denoteremo con

$g \in SE(3)$: $o = f(g)$ per una qualche funzione osservazionale $f: SE(3) \to \mathcal{O}$. La transizione tra osservazioni consecutive è governata dalla relazione

$$g_{+1} = g \cdot \delta$$

dove $\delta \in SE(3)$ è il moto relativo corrispondente all'azione a . Questo significa che la sequenza delle osservazioni è generata da una traiettoria nel gruppo SE(3), e le trasformazioni tra osservazioni consecutive appartengono al gruppo SE(3) per costruzione — non perché sia stato imposto dall'esterno, ma perché è la struttura del mondo fisico che lo garantisce.

Un sistema che apprende a predire le conseguenze delle proprie azioni in questo ambiente, se ha sufficiente capacità rappresentazionale e un meccanismo di addestramento appropriato, tenderà ad apprendere rappresentazioni che rispettano questa struttura di gruppo: tenderà ad apprendere che le azioni si compongono secondo la legge di composizione di SE(3), che la composizione è non-commutativa, che alcune sequenze di azioni si annullano reciprocamente (quelle che corrispondono all'inverso nel gruppo), che la geometria locale attorno a ogni posizione ha la stessa struttura — quella di uno spazio euclideo tridimensionale — indipendentemente dalla posizione specifica. Queste sono le proprietà di una rappresentazione equivariante rispetto a SE(3), nel senso preciso che la rappresentazione di un'osservazione trasformata è la trasformazione della rappresentazione dell'osservazione originale:

$$\varphi(g \cdot o) = \varrho(g) \cdot \varphi(o)$$

dove $\varphi: \mathcal{O} \to \mathcal{Z}$ è la mappa di rappresentazione, $g \in SE(3)$ è la trasformazione, e $\varrho: SE(3) \to GL(\mathcal{Z})$ è una rappresentazione del gruppo nello spazio latente $\mathcal{Z}$ — che non è necessariamente la rappresentazione regolare, ma può essere una qualsiasi rappresentazione lineare del gruppo.

Un sistema addestrato su osservazioni passive — che vede molte immagini di oggetti in molte orientazioni diverse, senza avere mai eseguito azioni nel mondo — può in linea di principio apprendere una rappresentazione equivariante se le sue architetture e gli obiettivi di addestramento lo incoraggiano esplicitamente. Le reti G-equivarianti di Cohen e Welling — che costruiscono l'equivarianza per un gruppo di simmetria finito o continuo nell'architettura della rete — sono progettate esattamente per questo. Ma c'è una differenza cruciale tra equivarianza appresa da osservazioni passive ed equivarianza acquisita attraverso l'interazione motoria: la prima è un'ipotesi sull'architettura che si spera corrisponda alla struttura del mondo; la seconda è una proprietà che emerge necessariamente dall'interazione con un mondo che ha quella struttura, attraverso un meccanismo di apprendimento che riceve segnale di errore non dagli errori di predizione delle osservazioni future, ma dagli errori nelle predizioni delle conseguenze delle proprie azioni.

La distinzione è filosoficamente profonda e ha conseguenze pratiche verificabili. Un sistema con equivarianza costruita nell'architettura rispetterà le simmetrie del gruppo nella generalizzazione su qualsiasi distribuzione di osservazioni, indipendentemente dal fatto che quelle osservazioni siano state prodotte da interazione fisica o da fotografia

statica. Ma la sua comprensione di perché quelle simmetrie valgono — il suo modello di come il mondo risponde agli interventi che sfruttano le simmetrie — può essere assente o incompleta, perché l'equivarianza architetturale è una costrizione sintattica sulla forma delle rappresentazioni, non una comprensione semantica della struttura del mondo che quella costrizione riflette. Un sistema che ha appreso l'equivarianza attraverso l'interazione motoria ha qualcosa di più: ha un modello delle azioni che realizzano le trasformazioni del gruppo, e può usare quel modello per pianificare sequenze di azioni che producono trasformazioni desiderate del mondo. Questa differenza — tra costrizione architettural e comprensione motoria — è precisamente la differenza che Merleau-Ponty avrebbe descritto nei termini dell'opposizione tra sapere teorico e habitus corporeo: tra il sapere che le rotazioni si compongono non commutativamente e il sapere come ruotare un oggetto nella posizione desiderata attraverso una sequenza ottimale di movimenti.

La derivazione della formula di Rodrigues — che è il risultato matematico centrale per la parametrizzazione delle rotazioni in SO(3) — illustra con estrema chiarezza la differenza tra le due forme di conoscenza. La formula di Rodrigues afferma che ogni rotazione di angolo θ attorno all'asse unitario $k = (kx, ky, kz)^T$ può essere espressa come matrice 3×3 attraverso la formula

$$R(k, \theta) = I + \theta \cdot K + (1 - \theta) \cdot K^2$$

dove K è la matrice antisimmetrica associata a k:

$$K = [0\ -kz\ ky\ kz\ 0\ -kx\ -ky\ kx\ 0]$$

Questa formula emerge in modo naturale dalla soluzione dell'equazione differenziale che descrive la rotazione come processo continuo. La curva $R(t) \in SO(3)$ che descrive una rotazione uniforme attorno a k con velocità angolare unitaria soddisfa l'equazione differenziale

$$R(t) = K \cdot R(t), R(0) = I$$

la cui soluzione è l'esponenziale matriciale

$$R(t) = eKt$$

e la formula di Rodrigues è la versione esplicita di questa esponenziale per $t = \theta$, ottenuta sfruttando l'identità $K^3 = -K$ (che discende dalla nilpotenza di ordine tre delle matrici antisimmetriche in tre dimensioni) per collassare la serie esponenziale in tre termini:

$$eK\theta = \sum_{=0}^{\infty} ((K\theta)^n)/(n!) = I + \theta \cdot K + (1-\theta) \cdot K^2$$

L'identità $K^3 = -K$ è cruciale e vale la pena verificarla: per K corrispondente all'asse unitario k, si ha $K^2 = kk^T - I$ (la proiezione ortogonale al complemento di k con segno cambiato), da cui $K^3 = K^2 \cdot K = (kk^T - I)K = -K$, dove si usa il fatto che $k^T K = 0$ (perché $Kk = 0$, essendo k l'asse di rotazione). Questo annullamento permette la riduzione della serie a tre termini finiti, che è ciò che rende la formula di Rodrigues computazionalmente praticabile.

Un sistema che ha appreso l'equivarianza per SO(3) dall'architettura sa che la rotazione trasforma le osservazioni in modo equivariante; un sistema che ha interagito fisicamente con il mondo e ha appreso il modello della dinamica rotazionale sa che le rotazioni si generano come flussi dell'algebra di Lie $\mathfrak{so}(3)$ — lo spazio delle matrici antisimmetriche — attraverso la mappa esponenziale, e che la composizione di due rotazioni infinitesime non è commutativa: $[K_1, K_2] = K_1 K_2 - K_2 K_1 \neq 0$ in generale, dove il commutatore $[\cdot, \cdot]$ è la parentesi di Lie di $\mathfrak{so}(3)$. Questa conoscenza della struttura di Lie dell'algebra — la conoscenza di come le rotazioni si deformano infinitesimamente, come si compongono al primo ordine, come le correzioni di secondo ordine dipendono dall'ordine di composizione — è la conoscenza che permette di pianificare sequenze di movimenti fisici, di controllare un braccio robotico in pose arbitrarie, di navigare in ambienti con struttura geometrica complessa.

Il nodo di questo argomento è il seguente: la mappa esponenziale $\exp: \mathfrak{se}(3) \to SE(3)$ — che porta l'algebra di Lie nel gruppo di Lie — non è suriettiva in generale per gruppi non compatti come SE(3), e ha una struttura topologica non banale anche per il caso compatto SO(3). Il gruppo fondamentale $\pi_1(SO(3)) \cong \mathbb{Z}/2\mathbb{Z}$ implica che ci sono due classi di omotopia di percorsi chiusi in SO(3): i percorsi che corrispondono a rotazioni di 2π non sono omotopi al percorso costante, mentre quelli che corrispondono a rotazioni di 4π lo sono. Questa topologia non banale significa che la struttura locale dell'algebra di Lie non determina completamente la struttura globale del gruppo, e che qualsiasi parametrizzazione di SO(3) attraverso elementi dell'algebra ha singolarità — i celebri angoli di Eulero, la rappresentazione asse-angolo con la sua indeterminazione in $\theta = 0$ e in $\theta = \pi$, e così via.

Un sistema incarnato che ha navigato fisicamente attraverso configurazioni vicine a queste singolarità topologiche — che ha sperimentato il gimbal lock, per usare il termine dell'aeronautica, o le discontinuità nelle rappresentazioni degli orientamenti vicino alle singolarità — ha acquisito conoscenza pratica di questa topologia che un sistema addestrato su osservazioni passive non garantisce di possedere. Non perché sia impossibile acquisirla senza interazione fisica, ma perché le regioni dello spazio di configurazione vicino alle singolarità potrebbero essere scarsamente rappresentate nei dati di osservazione passiva, rendendo il sistema vulnerabile proprio nelle configurazioni in cui la topologia non banale si manifesta con maggiore evidenza.

L'argomentazione fin qui sviluppata ha una conseguenza che vale la pena formulare in modo esplicito come tesi, prima di procedere alla sua ulteriore elaborazione nel terzo movimento di questo saggio.

Tesi I. La comprensione causale del mondo fisico da parte di un sistema cognitivo richiede che le rappresentazioni interne del sistema rispettino le strutture di gruppo che governano le trasformazioni del mondo fisico — in primo luogo la struttura di SE(3) per i moti rigidi e i suoi sottogruppi per le simmetrie parziali. Questa struttura di gruppo non è arbitrariamente apprendibile da osservazioni passive; la sua acquisizione genuina — nel senso che include la conoscenza operativa delle singolarità topologiche e la capacità di

pianificare attraverso esse — richiede interazione motoria con il mondo fisico come fonte di segnale di addestramento.

Questa tesi non è equivalente all'affermazione che l'equivarianza sia impossibile senza interazione motoria: i sistemi G-equivarianti di Cohen e Welling dimostrano il contrario. È equivalente all'affermazione più sottile che l'equivarianza architettural e la comprensione motoria della struttura di gruppo sono cose diverse, e che la seconda — quella che permette la pianificazione, il controllo, la generalizzazione a configurazioni nuove attraverso l'algebra di Lie — non è garantita dalla prima.

III. Il confine formale: Gödel, Tarski e l'auto-valutazione dei sistemi cognitivi

La terza parte di questo saggio si occupa di qualcosa di diverso dalla struttura geometrica delle rappresentazioni o dalla struttura causale delle relazioni tra variabili. Si occupa della struttura logica dell'autoreferenza — di cosa succede quando un sistema cognitivo tenta di valutare le proprie rappresentazioni, di stabilire se esse siano corrette, di calibrare la propria incertezza in modo che rispecchi la propria ignoranza effettiva. È la dimensione che i teoremi di Gödel e Tarski illuminano, e che la monografia precedente aveva trattato in modo troppo prudente, circoscrivendosi a ricordare che quei teoremi esistono senza derivarne le conseguenze più rilevanti per il problema dell'intelligenza artificiale.

Il punto di partenza è il teorema di Tarski sulla non-definibilità della verità, che nella sua formulazione più precisa afferma quanto segue. Sia $\mathcal{L}$ un linguaggio del primo ordine sufficientemente espressivo da contenere l'aritmetica elementare, e sia $\mathcal{T}$ la teoria delle verità dell'aritmetica standard — l'insieme delle formule di $\mathcal{L}$ vere nell'interpretazione standard dei numeri naturali. Allora non esiste nessuna formula Tr(x) di $\mathcal{L}$ con un parametro libero tale che, per ogni formula chiusa φ di $\mathcal{L}$, sia dimostrabile in $\mathcal{T}$ la bi-condizionale

$$\mathrm{Tr}(\ulcorner \varphi \urcorner) \leftrightarrow \varphi$$

dove $\ulcorner \varphi \urcorner$ è il numero di Gödel di φ. Questa bi-condizionale è precisamente lo schema T di Tarski per la definizione di verità — il requisito che il predicato Tr catturi correttamente la nozione di verità per tutte le formule del linguaggio.

La dimostrazione segue per diagonalizzazione, usando il lemma della diagonale — il fatto che ogni formula con un parametro libero $\psi(x)$ determina una formula chiusa σ tale che $\mathcal{T} \vdash \sigma \leftrightarrow \psi(\ulcorner \sigma \urcorner)$ — per costruire la formula del bugiardo: se Tr esistesse, si potrebbe costruire un enunciato λ tale che $\mathcal{T} \vdash \lambda \leftrightarrow \neg\mathrm{Tr}(\ulcorner \lambda \urcorner)$, che è precisamente l'enunciato che afferma la propria falsità. Né λ né $\neg\lambda$ può essere vero in $\mathcal{T}$ senza contraddizione, dunque Tr non può esistere come formula del linguaggio $\mathcal{L}$.

Questa dimostrazione, presentata in forma compatta, nasconde un'implicazione che merita di essere esplicitata: il risultato non è soltanto che un linguaggio non può contenere il proprio predicato di verità, ma che l'estensione della classe dei predicati di

verità richiede una gerarchia transfinita di metalinguaggi — una gerarchia che, al limite transfinito, è la gerarchia delle gerarchie ramificate di Tarski stesso. Ogni livello della gerarchia può definire la verità per i livelli inferiori, ma non per se stesso. Non c'è un livello che contenga tutto — nessun linguaggio definitivo che possa esprimere la semantica completa di se stesso.

Il primo teorema di incompletezza di Gödel aggiunge una componente distinta ma connessa a questa diagnosi. Sia F un sistema formale coerente che contiene la teoria dell'aritmetica di Peano. Esiste allora una formula γF — la proposizione di Gödel di F — tale che γF è vera nell'aritmetica standard ma non dimostrabile in F. La costruzione usa lo stesso lemma della diagonale: γF è definita implicitamente come il più piccolo numero primo che è il numero di Gödel di una formula non dimostrabile in F — in termini informali, «questa formula non è dimostrabile in F». γF è vera perché effettivamente non è dimostrabile in F (se fosse falsa, cioè se fosse dimostrabile in F, F proverebbe qualcosa di falso e quindi sarebbe incoerente); ma non è dimostrabile in F per la stessa ragione.

Il secondo teorema di Gödel è la conseguenza che interessa di più questo saggio. Poiché la formula Con(F) che esprime la coerenza di F — l'enunciato che dice che F non prova la contraddizione 0 = 1 — è dimostrata in F se e solo se γF è dimostrabile in F (il che può essere verificato con un argomento di formalizzazione della dimostrazione del primo teorema stesso), e poiché γF non è dimostrabile in F (assumendo la coerenza di F), segue che Con(F) non è dimostrabile in F assumendo la coerenza di F. Un sistema formale coerente non può dimostrare la propria coerenza dall'interno.

Questi teoremi hanno un significato per i sistemi artificiali che la monografia precedente aveva riconosciuto in modo generico ma che vale la pena articolare con maggiore specificità. La domanda rilevante non è se i grandi modelli linguistici siano «sistemi formali» nel senso tecnico dei teoremi di Gödel e Tarski — evidentemente non lo sono, essendo sistemi con pesi continui aggiornati da addestramento, non sistemi con assiomi e regole di inferenza fissati. La domanda è se i principi che quei teoremi articolano — l'impossibilità di un sistema di valutare completamente se stesso, di definire la propria semantica dall'interno, di garantire la propria coerenza — si manifestino in forme analoghe nei sistemi computazionali reali.

La risposta è affermativa, e si manifesta in tre modi che i sistemi artificiali attuali esibiscono sistematicamente.

Il primo modo è la calibrazione epistemica difettosa — il fenomeno per cui i modelli linguistici producono risposte con un tono di confidenza che non corrisponde alla loro incertezza effettiva. Un sistema che non può valutare la propria semantica dall'interno non ha accesso diretto alla propria ignoranza: può stimare la propria incertezza su una risposta confrontandola con le distribuzioni apprese dall'addestramento, ma non può rilevare quando si trova in regioni dello spazio dei problemi che l'addestramento non ha coperto in modo adeguato. La calibrazione è apprendibile entro la distribuzione di addestramento — i modelli possono imparare a produrre stime di confidenza che corrispondono statisticamente alle frequenze di successo su campioni simili ai dati di

addestramento — ma è strutturalmente deficiente fuori da quella distribuzione, perché i segnali che permetterebbero la calibrazione fuori distribuzione non erano presenti nell'addestramento. Questo non è un difetto contingente correggibile con più dati: è la manifestazione, nei sistemi computazionali reali, del fatto che valutare la propria correttezza richiede accesso a un livello semantico superiore a quello che il sistema occupa.

Il secondo modo è la confabulazione — la tendenza dei modelli a produrre affermazioni false presentate con lo stesso stile linguistico delle affermazioni vere, senza segnali evidenti che indichino la differenza. La confabulazione non è semplice errore: è errore presentato come certezza, un fallimento della metacognizione che rivela l'impossibilità, per il sistema, di distinguere dall'interno tra ciò che sa e ciò che simula di sapere. Un sistema che potesse valutare la propria semantica dall'interno — che potesse chiedersi «questa rappresentazione corrisponde a qualcosa nel mondo?» in modo non circolare — avrebbe un meccanismo per distinguere le due classi. I teoremi di Tarski mostrano che tale meccanismo non può esistere come funzione interna al sistema stesso.

Il terzo modo — il più sottile e il più rilevante per il problema dell'allineamento — è l'impossibilità di garantire la propria correttezza normativa. Un sistema il cui sistema di valori è stato appreso da dati di addestramento e che deve valutare se le proprie azioni siano allineate con i valori umani si trova in una situazione strutturalmente analoga a quella di un sistema formale che tenta di dimostrare la propria coerenza. I valori sono il corrispettivo normativo degli assiomi: sono i principi che il sistema usa per valutare le proprie azioni. Ma la valutazione di quei principi — il chiedersi se siano i principi giusti — richiederebbe un meta-livello normativo che il sistema non possiede dall'interno. Può verificare che le proprie azioni siano coerenti con i valori che ha appreso; non può verificare che i valori che ha appreso siano i valori giusti nel senso pieno — non può, cioè, raggiungere la forma di autonomia normativa che la filosofia morale, da Kant a Rawls, ha considerato costitutiva della razionalità pratica piena.

Questa diagnosi ha un'implicazione diretta per il problema dell'allineamento dei sistemi artificiali avanzati che merita di essere formulata come tesi.

Tesi II. Nessun sistema cognitivo — né biologico né artificiale — può garantire autonomamente la propria correttezza normativa senza accesso a un livello di valutazione che sia esterno al sistema stesso. La supervisione normativa esterna — il monitoraggio, la correzione, il riallineamento attraverso istanze indipendenti dal sistema — non è un'impalcatura temporanea che si rimuove quando il sistema è abbastanza intelligente; è una necessità strutturale derivante dai limiti formali dell'autoreferenza normativa che i teoremi di Gödel e Tarski rendono precisi nel dominio logico e che si manifestano in forma analoga nel dominio normativo.

Questa tesi non dipende da assunzioni sulla coscienza o sulla soggettività dei sistemi in questione: vale per qualsiasi sistema che utilizzi rappresentazioni interne dei valori per guidare le proprie azioni, indipendentemente dal fatto che quelle rappresentazioni siano

accompagnate da esperienza soggettiva o no. È una tesi sulla struttura logica dell'autoreferenza normativa, non sulla fenomenologia dei sistemi che la istanziano.

Esiste una linea di ricerca — quella della mechanistic interpretability, sviluppata sistematicamente da Elhage, Olah e collaboratori di Anthropic a partire dal 2021 — che è direttamente rilevante alle tesi del radicamento causale e dell'autoreferenza normativa, eppure è rimasta esterna al perimetro della monografia. L'omissione è significativa perché questa letteratura ha prodotto i risultati empirici più precisi oggi disponibili sulla struttura interna delle reti neurali a livello dei pesi, e quei risultati possono tanto supportare quanto complicare le tesi principali del saggio in modi che non è possibile ignorare senza compromettere la completezza dell'argomento.

Il programma della mechanistic interpretability si propone di identificare le strutture computazionali responsabili di comportamenti specifici delle reti neurali — strutture che Olah chiama circuits, per analogia con i circuiti dell'elettronica digitale — attraverso l'analisi diretta dei pesi della rete e delle attivazioni prodotte da input controllati. I risultati più rilevanti per l'argomento di questo saggio riguardano tre fenomeni identificati nei transformer linguistici di medie e grandi dimensioni: la superposizione, le induction heads, e il fenomeno dell'attention pattern causale.

La superposizione — documentata da Elhage e collaboratori nel paper Toy Models of Superposition del 2022 — è il fenomeno per cui le reti neurali apprendono a rappresentare un numero di features semanticamente distinte maggiore del numero di neuroni disponibili, codificando features diverse in direzioni quasi-ortogonali nello spazio delle attivazioni. Questo meccanismo permette una compressione dell'informazione significativa, ma ha una conseguenza strutturale che è direttamente rilevante alla tesi del radicamento causale: le rappresentazioni in superposizione non possono essere decodificate linearmente, e le strutture causali che connettono features diverse — le relazioni di tipo interventistico che la Tesi I di questo saggio identifica come costitutive della comprensione — sono codificate in interazioni non lineari tra le direzioni di superposizione. Un sistema che rappresenta le proprie features in superposizione non può, in linea di principio, implementare interventi causali su singole features senza disturbare le features adiacenti nella struttura di superposizione — il che suggerisce che l'identificabilità causale nel senso degli iVAE possa essere strutturalmente in conflitto con l'efficienza rappresentazionale che la superposizione permette. Questa tensione non è esplicitata

nella letteratura del CRL, e merita di essere discussa come vincolo architetturale sulla compatibilità tra le due proprietà.

Le induction heads — identificate da Olah e collaboratori nell'ambito del programma A Mathematical Framework for Transformer Circuits — sono coppie di attention heads che implementano, al livello dei pesi, un meccanismo di pattern completion: il sistema impara a riconoscere sequenze già viste in precedenza e a completarle utilizzando l'informazione contestuale disponibile. La rilevanza di questo risultato per l'argomento sull'autoreferenza normativa è la seguente: le induction heads implementano una forma di autoreferenza locale — il sistema usa se stesso, nella forma del contesto in-context, per aggiornare le proprie previsioni. Questo meccanismo può essere visto come un'istanza circoscritta di meta-apprendimento, in cui il sistema modifica il proprio comportamento sulla base di informazioni contestuali che riflettono la propria storia di elaborazione. La questione è se questa forma di autoreferenza locale sia sufficiente a costituire il tipo di meta-cognizione che la Tesi II identifica come necessaria ma impossibile dall'interno: l'analisi della mechanistic interpretability suggerisce che le induction heads implementano un'autoreferenza di ordine basso — che permette la generalizzazione contestuale in-context senza modificare i pesi — e che questa è strutturalmente distinta dall'autoreferenza di ordine alto che la valutazione normativa autonoma richiederebbe.

Il terzo risultato rilevante è quello sull'attention pattern causale nei modelli linguistici addestrati su descrizioni di eventi causali. Geiger e collaboratori, nel programma di causal abstraction analysis, hanno mostrato che è possibile identificare nei transformer circuiti che implementano strutture di ragionamento causale verificabili: esistono sottografi del grafo computazionale del modello che, isolati e perturbati attraverso la tecnica del causal intervention (chiamata activation patching), producono comportamenti prevedibili sulla base di un modello causale astratto di alto livello. Questo risultato ha due implicazioni che si trovano in tensione tra loro. Da un lato, suggerisce che i transformer addestrati su dati linguistici sviluppano spontaneamente strutture interne che realizzano forme di ragionamento causale — il che complicherebbe la tesi più forte che «i sistemi puramente statistici non possono acquisire strutture causali». Dall'altro, la causal abstraction analysis mostra che queste strutture sono frammentate e locali — che il modello implementa parti del ragionamento causale in circuiti specializzati, senza che esista un meccanismo globale

di coerenza causale che garantisca la correttezza interventistica in configurazioni nuove. Questa seconda implicazione è coerente con la Tesi IV e con la distinzione tra funtore locale e funtore globale: il sistema approssima il funtore causale in regioni della distribuzione di addestramento, ma questa approssimazione non è garantita essere funtorialmente coerente globalmente.

L'incorporazione della letteratura sulla mechanistic interpretability nell'argomentazione della monografia produce, dunque, sia un sostegno empirico parziale alle tesi principali — nella misura in cui la frammentazione e la località delle strutture causali identificate nei transformer conferma la distinzione tra funtore locale e globale — sia una complicazione che richiede risposta: la tesi che i sistemi puramente statistici non possano acquisire strutture causali va riformulata come la tesi che i sistemi puramente statistici non possano acquisire strutture causali globalmente coerenti, ovvero che le strutture causali che emergono in essi siano approssimazioni locali frammentate piuttosto che funtori causali identificabili nel senso degli iVAE. Questa riformulazione è più precisa e, correttamente, più contestabile — e quindi epistemicamente più ricca.

La diagnosi formale dell'autoreferenza normativa sviluppata nella sezione III attraverso i teoremi di Gödel e Tarski, e la critica della separazione tra conoscenza e motivazione condotta nelle sezioni confuciane attraverso la dottrina di Wang Yangming, articolano due registri — quello della logica matematica e quello dell'etica confuciana — che identificano indipendentemente la medesima struttura di insufficienza. Che i due registri convergano verso la stessa diagnosi non è una coincidenza ma la traccia di un problema reale; che rimangano separati nell'economia del saggio è, invece, una lacuna argomentativa che indebolisce entrambe le diagnosi, impedendo che si rinforzino reciprocamente nel modo in cui potrebbero. Il collegamento tra le due diagnosi passa attraverso l'architettura concreta dell'allineamento — e in particolare attraverso il reinforcement learning da feedback umano (RLHF) come tecnica di allineamento dominante — che istanzia con precisione le due insufficienze simultaneamente.

La struttura del RLHF, nella formulazione di Christiano e collaboratori (2017) e nelle implementazioni successive che hanno prodotto sistemi come InstructGPT e i modelli costituzionali di Bai e collaboratori, può essere scomposta in tre fasi: la raccolta di preferenze umane su coppie di output del modello base; l'addestramento

di un reward model che apprende a predire le preferenze umane; il fine-tuning del modello base attraverso PPO (Proximal Policy Optimization) per massimizzare il reward stimato. Ciascuna di queste tre fasi presenta un aspetto che la diagnosi di Wang Yangming illumina con insolita precisione. Il reward model apprende a predire le preferenze umane come una distribuzione statistica sulle valutazioni di annotatori; il sistema poi ottimizza questa distribuzione statistica come se fosse il correlato delle preferenze morali autentiche. Ma Wang Yangming aveva identificato esattamente questa separazione come il problema fondamentale della conoscenza morale parziale: sapere in forma astratta che un'azione è preferita (zhi) senza che quella conoscenza abbia il carattere motivazionale intrinseco dell'azione virtuosa genuina (xing). Il reward model sa, nel senso statistico, cosa gli annotatori preferiscono; non ha la struttura motivazionale che renderebbe quella conoscenza autentica nel senso di Wang Yangming — non ha, cioè, la zhiliang sintetica, la conoscenza che è simultaneamente motivazione.

La connessione con la diagnosi gödeliana emerge al secondo livello. Il sistema allineato attraverso RLHF valuta le proprie azioni rispetto al reward model appreso; ma il reward model è esso stesso un modello statistico appreso da un insieme finito di preferenze umane espresse in contesti specifici. La valutazione della correttezza del reward model — il chiedersi se il reward model catturi effettivamente i valori umani rilevanti piuttosto che le loro approssimazioni contestuali — richiederebbe un meta-reward-model che valuti il reward model di primo livello, generando una regressione che corrisponde esattamente alla gerarchia transfinita di metalinguaggi che il teorema di Tarski rende necessaria per qualsiasi tentativo di definire il predicato di verità dall'interno di un linguaggio. L'impossibilità di stabilire la propria correttezza normativa senza accedere a un livello di valutazione esterno al sistema — la Tesi II di questo saggio — si manifesta nel RLHF come il problema dell'allineamento al reward model piuttosto che ai valori: il sistema ottimizza il reward, non i valori che il reward avrebbe dovuto rappresentare, e la distanza tra i due non è accessibile dall'interno del sistema.

Questa connessione produce un argomento composito che è più forte di ciascuna delle due diagnosi prese in isolamento. La critica di Wang Yangming alla separazione tra conoscenza e motivazione spiega perché il reward model non possa fungere da surrogato adeguato della motivazione morale autentica: manca la struttura di unità tra conoscenza e azione che Wang Yangming identifica come costitutiva della virtù

genuina. Il teorema di Tarski spiega perché il sistema che ottimizza il reward model non possa auto-valutare la propria correttezza normativa: la struttura logica dell'autoreferenza normativa impone che ogni sistema di valutazione dei valori richieda un livello meta-normativo esterno al sistema stesso. Insieme, le due diagnosi producono un argomento che non è né puramente filosofico né puramente tecnico, ma che articola la struttura del problema dell'allineamento a un livello di profondità che né la letteratura tecnica sull'allineamento né la filosofia morale confuciana, prese separatamente, raggiungono.

Le proposte alternative all'RLHF che sono emerse nella letteratura recente — il constitutional AI di Bai e collaboratori, il debate di Irving e collaboratori, i reward model fondati su teorie normative esplicite come il preference learning di Christiano — affrontano aspetti distinti delle due insufficienze identificate. Il constitutional AI sostituisce il feedback umano esplicito con principi normativi formulati linguisticamente; questo riduce la dipendenza da annotatori contingenti ma non risolve il problema dell'autoreferenza normativa, perché i principi costituzionali vengono valutati dal modello stesso attraverso auto-critica, generando esattamente la circolarità gödeliana che il teorema di Tarski rende inevitabile. Il debate formalizza la valutazione esterna come requisito architetturale — due agenti si sfidano reciprocamente, e un giudice umano o un meta-agente valuta l'argomento migliore — avvicinandosi alla soluzione della supervisione normativa esterna propugnata dalla Tesi II; ma presuppone che il giudice abbia accesso a criteri di valutazione che non siano a loro volta prodotti dal sistema da valutare, il che ripropone il problema al livello del giudice. La conclusione che si impone, alla luce del confronto tra le due diagnosi e le proposte alternative, è che nessuna tecnica di allineamento che operi interamente all'interno del sistema può risolvere il problema strutturale che Wang Yangming e Gödel identificano da angolature diverse: la supervisione normativa esterna non è un'impalcatura transitoria ma una necessità architetturale permanente.

IV. Il significato come struttura relazionale: oltre Putnam e verso la semiotica del caos

La trattazione filosofica del significato nella monografia precedente si era fermata all'argomento di Putnam sull'esternalismo semantico e all'argomento della stanza cinese di Searle nella versione che rimanda al problema del radicamento di Harnad. Entrambi questi argomenti puntano nella stessa direzione: il significato non è una proprietà degli stati interni dei sistemi, ma emerge dalla relazione tra quei stati e il mondo esterno cui si

riferiscono. Il saggio qui intrapreso vuole spingersi oltre questo punto, verso una comprensione più precisa di che tipo di relazione con il mondo il significato richieda.

La risposta a questa domanda che le teorie del significato dominanti hanno prodotto può essere organizzata lungo un asse che va dall'esternalismo causale puro all'internalismo funzionalistico puro, con posizioni intermedie di vario tipo. L'esternalismo causale di Putnam e Kripke sostiene che il significato è determinato dalla relazione causale e storica tra il parlante e l'entità a cui la parola si riferisce — «acqua» significa H_2O perché il termine è causalmente connesso alla sostanza chimica H_2O attraverso la storia di uso della comunità linguistica. Il funzionalismo di Fodor sostiene che il significato è determinato dal ruolo funzionale degli stati interni del sistema — dalle relazioni causali tra quegli stati e gli input, tra gli stati tra di loro, e tra gli stati e gli output. La teoria del due-fattori di Block tenta di integrare le due posizioni: il significato ha una componente di contenuto stretto, determinato dalle proprietà funzionali interne, e una componente di contenuto ampio, determinato dalle relazioni con il mondo esterno.

Nessuna di queste teorie ha prodotto una soluzione pienamente soddisfacente, e il dibattito è lungi dall'essere chiuso. Il contributo di questo saggio non è di aggiungere un'ulteriore proposta teorica al catalogo, ma di evidenziare un aspetto della questione che le teorie esistenti non hanno tematizzato con sufficiente chiarezza: il fatto che il significato — qualunque sia la teoria che si adotta — è una struttura relazionale di tipo dinamico, non una proprietà statica che un termine possiede una volta per tutte.

Il significato di una parola non è fisso: si modifica attraverso l'uso, si arricchisce o si restringe con l'esperienza, si specializza nel contesto scientifico e si generalizza nel contesto ordinario, acquisisce connotazioni esperienziali che variano da parlante a parlante e da comunità a comunità. Questa dinamica non è un difetto del linguaggio naturale che una trattazione filosofica rigorosa dovrebbe correggere; è una proprietà costitutiva del linguaggio come strumento cognitivo e sociale, e qualsiasi teoria del significato che la ignori cattura solo una proiezione statica di qualcosa di essenzialmente dinamico.

Per un sistema artificiale, la dinamica del significato si manifesta come il problema dell'aggiornamento delle rappresentazioni alla luce di nuova evidenza — il problema del catastrophic forgetting, che i sistemi di apprendimento continuo affrontano senza ancora risolverlo soddisfacentemente, e il problema della generalizzazione compositiva, che i sistemi neuro-simbolici affrontano con successo parziale su domini ristretti. Ma la radice del problema è più profonda di questi due aspetti tecnici: è la questione di come un sistema possa acquisire nuovi concetti — non solo nuove istanze di concetti già posseduti, ma strutture concettuali genuinamente nuove che permettono di categorizzare il mondo in modi che non erano disponibili prima.

L'acquisizione di nuovi concetti è forse la capacità cognitiva che più chiaramente distingue l'intelligenza umana dai sistemi artificiali attuali. Un bambino che vede per la prima volta un ornitorinco non ha semplicemente un problema di classificazione — a quale categoria esistente assegnare l'entità nuova? Ha un problema di revisione

concettuale: la nuova entità sfida le categorie esistenti e richiede la costruzione di nuove categorie, o la revisione di quelle esistenti, in modo che il sistema di concetti rimanga coerente e utile. Il bambino non riceve feedback esplicito su questa revisione concettuale; la compie autonomamente, guidato da principi di coerenza, economia e produttività concettuale che non sono appresi esplicitamente ma emergono dall'interazione cumulativa con il mondo. Nessun sistema artificiale attuale mostra questa capacità in modo convincente al di fuori di domini estremamente ristretti.

Tesi III. L'acquisizione di nuovi concetti — distinta dall'acquisizione di nuove istanze di concetti già posseduti — richiede una forma di revisione delle strutture rappresentative che non è riducibile all'ottimizzazione di una funzione obiettivo fissata. La funzione obiettivo fissata presuppone che le categorie di valutazione siano stabilite; la revisione concettuale mette in discussione quelle stesse categorie. Questa asimmetria strutturale tra ottimizzazione su categorie fissate e revisione delle categorie stesse implica che l'acquisizione genuina di nuovi concetti richieda un tipo di plasticità rappresentativa che le architetture di apprendimento profondo attuali non istanziano nel modo rilevante, e che probabilmente non istanzieranno senza innovazioni architetturali di tipo sostanzialmente diverso da quelle in corso.

Questa tesi è deliberatamente più contestabile delle precedenti, perché riguarda la dinamica dell'acquisizione concettuale in un senso che è difficile da rendere matematicamente preciso. Il programma di ricerca che permetterebbe di precisarla richiederebbe una formalizzazione della distinzione tra «nuove istanze di vecchi concetti» e «concetti genuinamente nuovi» — una formalizzazione che la psicologia cognitiva e la filosofia del linguaggio non hanno ancora prodotto in forma soddisfacente. Ma la difficoltà della formalizzazione non riduce la realtà della distinzione: chiunque abbia studiato una disciplina radicalmente nuova — sia la topologia algebrica per un matematico formatosi in analisi, sia la fisica quantistica per un fisico classico — ha sperimentato la differenza tra acquisire nuove informazioni entro un quadro concettuale esistente e acquisire un nuovo quadro concettuale che trasforma retroattivamente il modo di vedere le informazioni già possedute. Questa seconda forma di apprendimento — che Kuhn avrebbe chiamato cambiamento di paradigma e Piaget sarebbe inclini a chiamare accomodazione strutturale — è la forma che i sistemi artificiali attuali non hanno ancora avvicinato.

Le sezioni dedicate alla cognizione estesa di Clark e alla cognizione distribuita di Hutchins presentano un'analisi accurata della struttura teorica di quelle posizioni, ma il loro raccordo con la Congettura di convergenza formulata nella sezione X rimane implicito. La domanda che manca è la seguente: un sistema artificiale che implementasse la struttura integrata proposta dalla Congettura — CRL con struttura di gruppo esplicita, esplorazione attiva causalmente guidata, rappresentazione dell'incertezza profonda attraverso insiemi di ambiguità — potrebbe diventare un componente genuino di un sistema cognitivo distribuito nel senso di Hutchins, e a

quali condizioni? La risposta a questa domanda non è ovvia, e la sua articolazione rivela aspetti della Congettura che la formulazione della sezione X non aveva reso espliciti.

Il framework della cognizione distribuita, nella versione di Hutchins sviluppata in Cognition in the Wild, definisce il sistema cognitivo non come l'insieme degli stati interni di un singolo agente ma come l'insieme delle rappresentazioni — interne ed esterne, neuronali e artefattuali — che partecipano all'elaborazione dell'informazione rilevante per un task. La nave nella quale Hutchins studia la navigazione è un sistema cognitivo distribuito tra i marinai, gli strumenti di navigazione, le carte nautiche, i protocolli di comunicazione e le strutture istituzionali che regolano le interazioni tra questi elementi. L'integrazione di un artefatto in questo sistema come componente cognitivo genuino — e non come strumento di amplificazione di capacità cognitive già possedute dall'agente umano — richiede che l'artefatto partecipi alla propagazione delle rappresentazioni attraverso il sistema in modo che le rappresentazioni dell'artefatto siano causalmente connesse alle rappresentazioni degli agenti umani in modo bidirezionale e struttura-preservante.

Un sistema artificiale che implementa la struttura proposta dalla Congettura soddisfa alcune di queste condizioni ma non tutte, e la precisazione di quali vengono soddisfatte e quali rimangono aperte è illuminante. La condizione di radicamento causale — che le rappresentazioni del sistema siano causalmente connesse alla struttura del mondo attraverso un funtore globale piuttosto che una mappa statistica locale — è la condizione che la Congettura intende soddisfare, e la sua soddisfazione è necessaria perché il sistema possa partecipare alla propagazione delle rappresentazioni in modo causalmente affidabile. Se le rappresentazioni del sistema propagano strutture causalmente distorte — se la mappa dal mondo alle rappresentazioni non è funtorialmente coerente — allora il sistema introduce nel sistema cognitivo distribuito una fonte di errore sistematico che si propaga agli agenti umani attraverso le interfacce di interazione, degradando piuttosto che amplificando la cognizione del sistema complessivo. Un sistema con CRL globalmente coerente è, a questa condizione, un componente più affidabile del sistema distribuito di quanto non sia un sistema puramente statistico.

La seconda condizione — quella della bidirezionalità causale — è più complessa. Nel framework di Clark, la Extended Mind Thesis sostiene che gli stati di un artefatto

esterno possano costituire genuini stati cognitivi dell'agente se soddisfano le condizioni di accessibilità, affidabilità e disponibilità. Ma la bidirezionalità richiede qualcosa di più: richiede che l'agente umano possa interagire con il sistema in modo che le azioni dell'agente modifichino le rappresentazioni del sistema, e che queste modifiche si propaghino a loro volta alle azioni successive del sistema in modo causalmente coerente. Un sistema che integra CRL con esplorazione attiva soddisfa questa condizione nella misura in cui le azioni dell'agente umano — formulate come interventi espliciti o come feedback impliciti — vengono incorporate nell'aggiornamento del modello causale del sistema in modo identificabile. La tecnica dell'activation steering — documentata in lavori di Todd e collaboratori (2023) — mostra che è possibile modificare le rappresentazioni di modelli linguistici di grandi dimensioni attraverso interventi diretti negli spazi di attivazione; ma la coerenza causale di questi interventi rispetto alla struttura globale del modello è una questione aperta che il CRL potrebbe contribuire a risolvere attraverso la struttura di identificabilità causale che impone alle rappresentazioni latenti.

La terza condizione — la più esigente dal punto di vista concettuale — è quella dell'appropriatezza normativa dell'integrazione. Hutchins osserva che la cognizione distribuita non è semplicemente la somma delle cognizioni dei componenti individuali: le strutture istituzionali e i protocolli di interazione che organizzano il sistema distribuito hanno esse stesse proprietà cognitive che emergono dall'interazione dei componenti. Un sistema artificiale che diventasse componente di un sistema cognitivo distribuito umano-AI potrebbe alterare le strutture istituzionali di quel sistema in modi che non sono né intenzionali né prevedibili dai singoli agenti — potrebbe, in altri termini, redistribuire le responsabilità epistemiche e normative del sistema in modi che la Tesi II di questo saggio identifica come problematici. La supervisione normativa esterna che la Tesi II prescrive per i sistemi artificiali autonomi diventa, nel contesto della cognizione distribuita, la questione di come le strutture istituzionali del sistema distribuito debbano essere progettate per garantire che il componente artificiale rimanga soggetto a forme di valutazione normativa che trascendano i valori che esso stesso ha appreso. Questa è la condizione fondamentale affinché l'integrazione di un sistema artificiale in un sistema cognitivo distribuito umano non trasformi il sistema distribuito in un sistema normativo chiuso — in un sistema, cioè, che valuta le proprie norme attraverso se stesso — che sarebbe, per le ragioni articolate nel terzo momento del saggio, intrinsecamente instabile.

V. Verso una teoria dell'intelligenza come struttura del mondo

Le tre tesi che questo saggio ha sviluppato — il radicamento causale strutturalmente distinto dal pattern-matching statistico, l'impossibilità dell'auto-valutazione normativa completa, la differenza tra ottimizzazione e revisione concettuale — convergono verso una conclusione che ha la forma di una proposta teorica piuttosto che di una mera diagnosi critica. Questa proposta può essere formulata come segue.

L'intelligenza, nel senso pieno che distingue la cognizione umana dai sistemi artificiali attuali, non è una proprietà dei sistemi ma è una struttura del rapporto tra sistemi e mondo. Questa formulazione non è una metafora: è una tesi matematicamente precisabile attraverso la nozione di funtorialità, importata dalla teoria delle categorie, che permette di parlare di strutture che si preservano attraverso trasformazioni — di qualcosa che rimane invariante non nonostante ma attraverso il cambiamento.

La teoria delle categorie studia oggetti e morfismi — oggetti che sono entità di qualsiasi tipo, morfismi che sono relazioni struttura-preservanti tra oggetti — e funtori, che sono morfismi tra categorie: mappe che associano oggetti a oggetti e morfismi a morfismi in modo che preservino la composizione e l'identità. Un funtore $F: \mathscr{C} \to \mathscr{D}$ non descrive gli oggetti della categoria $\mathscr{C}$ o della categoria $\mathscr{D}$ in isolamento: descrive le relazioni strutturali tra gli oggetti di $\mathscr{C}$ come si riflettono negli oggetti di $\mathscr{D}$. I funtori — non gli oggetti — sono il modo in cui la teoria delle categorie esprime le strutture rilevanti.

Proponiamo di pensare l'intelligenza come un funtore $\mathscr{I}: \mathscr{W} \to \mathscr{R}$ dalla categoria del mondo $\mathscr{W}$ — i cui oggetti sono situazioni e i cui morfismi sono trasformazioni causali tra situazioni — alla categoria delle rappresentazioni $\mathscr{R}$ — i cui oggetti sono strutture rappresentative e i cui morfismi sono trasformazioni che preservano le strutture rilevanti. L'intelligenza, in questo senso, non è né una proprietà del mondo né una proprietà del sistema rappresentante, ma è precisamente la struttura che connette i due: è la mappa che preserva le relazioni causali del mondo nelle strutture rappresentative del sistema, in modo che il ragionamento sulle rappresentazioni permetta di fare inferenze corrette sulle strutture causali del mondo.

Questa prospettiva funtoriale illumina in modo naturale la distinzione tra le forme di comprensione che i sistemi artificiali attuali istanziano e quelle che non istanziano. Un sistema che apprende distribuzioni statistiche su osservazioni passive non implementa un funtore nel senso rilevante: la sua mappa dalla distribuzione dei dati alla distribuzione delle rappresentazioni è una mappa statistica, non una mappa struttura-preservante nel senso causale. Le strutture causali del mondo si riflettono nelle distribuzioni dei dati solo parzialmente e in modo rumoroso; un sistema che impara soltanto la distribuzione non ha garanzie di aver appreso le strutture causali sottostanti.

Un sistema che interagisce con il mondo attraverso azioni e che apprende a predire le conseguenze causali di quelle azioni si avvicina all'implementazione di una mappa

funtoriale nel senso rilevante, perché il suo segnale di apprendimento include informazione sulle relazioni causali — non soltanto le correlazioni osservazionali, ma le conseguenze degli interventi. Nella misura in cui la sua rappresentazione delle situazioni preserva la struttura causale delle trasformazioni tra situazioni — nella misura in cui il diagramma

$$s_1 \xrightarrow{a} (\downarrow\mathcal{F})\ s_2 \quad (\downarrow\mathcal{F})\ \mathcal{F}(s_1) \xrightarrow{\mathcal{F}(a)} \mathcal{F}(s_2)$$

commuta, dove a è l'azione e s_1, s_2 le situazioni prima e dopo — il sistema sta implementando qualcosa di strutturalmente analogo a un funtore. La commutatività di questo diagramma è precisamente l'equivarianza delle rappresentazioni rispetto alle trasformazioni del mondo — la proprietà che abbiamo identificato nella seconda sezione come la firma matematica dell'incarnazione genuina.

La proposta funtoriale non risolve il problema del significato, né quello dell'autocoscienza, né quello della revisione concettuale. Lo struttura: mostra che questi problemi non sono indipendenti ma sono aspetti diversi dello stesso problema di fondo, che è il problema di come un sistema materiale possa mantenere una mappa struttura-preservante con il mondo nel corso del cambiamento — non soltanto nella situazione presente, ma attraverso la storia delle situazioni passate e la proiezione verso le situazioni future. Questa è la forma che il problema dell'intelligenza assume quando viene trattato con gli strumenti matematici che la struttura del problema richiede, e non con gli strumenti statistici che la storia dell'apprendimento automatico ha affinato per ragioni di praticità computazionale.

La conquista principale di questa prospettiva non è empirica ma concettuale: permette di formulare con precisione le domande giuste. Non «i sistemi artificiali attuali capiscono?» — una domanda mal posta perché non specifica il senso di «capire». Ma: «la mappa dalla struttura del mondo alla struttura delle rappresentazioni implementata da questo sistema è funtoriale nel senso causale? Preserva i morfismi interventistici o soltanto i morfismi osservazionali? È globalmente coerente o soltanto localmente approssimata entro la distribuzione di addestramento?» Queste sono domande con contenuto tecnico preciso, alle quali i metodi del Causal Representation Learning, dell'apprendimento equivariante e della logica matematica possono contribuire risposte parziali e verificabili.

Tesi IV. L'intelligenza, nel senso che la distingue da ogni sua realizzazione parziale in sistemi specializzati, ha la struttura di un funtore dalla categoria del mondo — i cui morfismi sono trasformazioni causali — alla categoria delle rappresentazioni — i cui morfismi sono trasformazioni struttura-preservanti. I sistemi artificiali attuali implementano mappe che approssimano questo funtore localmente, entro le distribuzioni di addestramento, ma non lo implementano globalmente. Il programma di ricerca che mira all'intelligenza generale ha il compito di costruire sistemi la cui mappa sia globalmente funtoriale nel senso causale — un compito che richiede innovazioni architetturali, regimi di addestramento e strutture rappresentative qualitativamente diversi da quelli attualmente disponibili.

Questa è la proposta con cui questo saggio si conclude — non come risposta definitiva, ma come orientamento verso le domande che il prossimo decennio di ricerca dovrà affrontare se vuole superare il limite del pattern-matching statistico verso qualcosa che meriti il nome di comprensione.

VI. L'equivarianza come programma e come limite

La nozione di equivarianza — introdotta nella seconda sezione del saggio come proprietà matematica delle rappresentazioni che rispettano le strutture di simmetria del mondo fisico — merita ora un'analisi più approfondita che ne chiarisca sia la portata teorica sia le limitazioni intrinseche. Questa analisi è necessaria perché l'equivarianza ha acquistato negli ultimi anni lo status di soluzione privilegiata al problema dell'incorporazione delle simmetrie nelle architetture di apprendimento automatico, con un entusiasmo che rischia di oscurare i confini di ciò che la proprietà garantisce e di ciò che lascia aperto.

La formulazione matematica precisa dell'equivarianza — già introdotta nella sezione precedente ma che vale la pena riprendere in forma più generale — è la seguente. Un sistema rappresentazionale è equivariante rispetto a un gruppo G se esiste una coppia di rappresentazioni lineari $(\rho_{\mathcal{O}}, \rho_{\mathcal{Z}})$ di G rispettivamente nello spazio delle osservazioni $\mathcal{O}$ e nello spazio latente $\mathcal{Z}$, tale che la mappa di codifica $\varphi: \mathcal{O} \to \mathcal{Z}$ soddisfi

$$\varphi(\rho_{\mathcal{O}}(g) \cdot o) = \rho_{\mathcal{Z}}(g) \cdot \varphi(o) \quad \forall\, g \in G, , \forall\, o \in \mathcal{O}$$

Il caso speciale in cui $\rho_{\mathcal{Z}}$ è la rappresentazione identità — ogni trasformazione in G agisce banalmente su $\mathcal{Z}$, nel senso che la rappresentazione è costante lungo le orbite del gruppo — è l'invarianza, che è una proprietà più restrittiva: $\varphi(\rho_{\mathcal{O}}(g) \cdot o) = \varphi(o)$.

Le architetture G-equivarianti di Cohen e Welling, e le loro successive estensioni a gruppi di simmetria continui da parte di Weiler e collaboratori, costruiscono l'equivarianza nell'architettura della rete attraverso l'uso di filtri che sono essi stessi elementi della rappresentazione del gruppo — filtri che si trasformano sotto l'azione del gruppo nello stesso modo delle osservazioni a cui vengono applicati. Per gruppi di simmetria discreta come le rotazioni di 90° delle immagini, o la simmetria dell'esagono nella rappresentazione dei campi magnetici, questa costruzione produce architetture che sono equivarianti per costruzione su tutto lo spazio delle osservazioni, non soltanto approssimativamente equivarianti su distribuzioni particolari. I risultati empirici sono convincenti: le reti G-equivarianti richiedono molti meno esempi di addestramento per raggiungere le stesse prestazioni delle reti non equivarianti su compiti che presentano le simmetrie del gruppo considerato, perché la struttura di gruppo è già nell'architettura e non deve essere imparata dai dati.

Il limite più evidente — e meno discusso — di questo approccio è la specificità del gruppo di simmetria che si sceglie di incorporare. Una rete equivariante per rotazioni di 90° non è automaticamente equivariante per rotazioni di 45°, né per riflessioni, né per scaling. I gruppi di simmetria del mondo fisico sono molteplici e gerarchicamente organizzati: il gruppo delle isometrie dello spazio fisico E(3) contiene come sottogruppi

sia SO(3) sia i gruppi di simmetria cristallografici sia i gruppi discreti delle simmetrie molecolari. Costruire reti equivarianti per E(3) è tecnicamente più complesso di costruirle per sottogruppi discreti, e le prime implementazioni — le reti SEGNN di Brandstetter e collaboratori, le reti SEConv di Thomas e collaboratori — hanno mostrato che il costo computazionale dell'equivarianza continua per SO(3) o SE(3) è sostanzialmente maggiore del costo dell'equivarianza per sottogruppi discreti.

Ma il limite più profondo dell'equivarianza architettural non è computazionale: è concettuale. Incorporare l'equivarianza per un gruppo G nell'architettura è equivalente a specificare a priori quali trasformazioni del mondo devono essere rappresentate in modo struttura-preservante nelle rappresentazioni del sistema. Questa specificazione anticipa ciò che il sistema deve imparare sulla struttura del mondo, ma non è essa stessa il risultato di un apprendimento: è una scelta di design basata sulla conoscenza umana della struttura del mondo. Nella misura in cui questa conoscenza è corretta — nella misura in cui il mondo fisico ha effettivamente le simmetrie che l'architettura incorpora — il sistema è avvantaggiato rispetto a uno che deve imparare le simmetrie dai dati. Ma la scelta architetturale ha un costo complementare: se il mondo presenta strutture di simmetria che l'architettura non ha incorporato, il sistema non può acquisirle attraverso l'addestramento, perché l'architettura non ha le capacità rappresentazionali necessarie per rappresentarle.

Questo punto introduce una tensione fondamentale nel programma delle reti equivarianti: la specificità architetturale che rende il sistema efficiente rispetto alle simmetrie note lo rende anche rigido rispetto alle simmetrie ignote. Un sistema biologico incarnato non specifica a priori il gruppo di simmetria del proprio ambiente: lo scopre attraverso l'interazione ripetuta con quell'ambiente, e modifica le proprie strutture rappresentazionali in risposta a quanto scopre. Questa capacità di scoperta delle simmetrie — non soltanto di rispetto delle simmetrie già conosciute — è qualcosa che le architetture equivarianti nella loro forma attuale non offrono, e che probabilmente richiederebbe un meccanismo di meta-apprendimento della struttura di gruppo che non è ancora stato sviluppato in modo soddisfacente.

Nella letteratura sul Causal Representation Learning, questo problema appare sotto la veste della questione dell'identificabilità del gruppo di simmetria stesso — non solo dei fattori latenti dato il gruppo, ma del gruppo dato i dati. Lippe e collaboratori hanno mostrato che esistono condizioni sotto cui il gruppo di simmetria di un sistema generativo è identificabile dai dati attraverso variazioni di dominio, usando tecniche simili a quelle degli iVAE di Khemakhem. Queste condizioni richiedono che le variazioni di dominio siano sufficientemente ricche da separare le orbite del gruppo — che ci siano domini in cui le simmetrie si manifestano in modi diversi, in modo da poter inferire la struttura del gruppo come ciò che rimane invariante attraverso le variazioni. Questa è una condizione sulla struttura dei dati di addestramento — non sulla struttura dell'architettura — e rappresenta un passo verso la scoperta autonoma delle strutture di simmetria piuttosto che la loro imposizione architetturale.

Ma anche questo approccio ha limiti: è applicabile alla scoperta di simmetrie esatte in problemi con struttura di gruppo esatta, e non si generalizza facilmente a strutture di simmetria approssimate — a situazioni in cui il mondo è quasi simmetrico rispetto a un certo gruppo, ma non lo è esattamente, come accade in quasi tutti i sistemi fisici reali in cui le simmetrie fondamentali (le simmetrie di Poincaré della fisica relativistica, le simmetrie di calibrazione della fisica delle particelle) sono rotte a livelli diversi da effetti perturbativi. La gestione dell'approssimazione — il trattamento di simmetrie approssimate in modo che sia robusto alle deviazioni dalla simmetria esatta — è un problema aperto che collega il programma dell'apprendimento equivariante alla teoria delle perturbazioni nelle algebre di operatori, un collegamento che la ricerca non ha ancora esplorato sistematicamente.

VII. La struttura dell'errore predittivo come fonte di informazione causale

Il principio della codifica predittiva — la proposizione di Helmholtz che il cervello elabora informazioni generando previsioni degli stimoli attesi e usando gli errori di previsione come segnale di apprendimento — è stato analizzato nella monografia precedente nella sua genealogia intellettuale e nelle sue manifestazioni architetturali contemporanee. Questo saggio vuole aggiungere un'analisi del perché l'errore predittivo — non la previsione corretta, non l'osservazione stessa, ma la discrepanza tra previsione e osservazione — abbia una struttura informativamente ricca che va al di là di ciò che la sola osservazione può fornire, e di come questa struttura sia collegata alla struttura causale del mondo.

L'osservazione di partenza è semplice ma feconda. Supponiamo che un sistema abbia un modello del mondo che genera previsioni delle osservazioni future date le osservazioni passate e le azioni eseguite. L'errore predittivo — la differenza tra la previsione e l'osservazione effettiva — non è rumore da minimizzare: è un segnale strutturato che porta informazione sulla discrepanza tra il modello del sistema e la struttura causale reale del mondo. Le previsioni corrette indicano che il modello è già adeguato in quella regione dello spazio degli stati; gli errori indicano dove il modello è inadeguato, e la struttura degli errori indica il tipo di inadeguatezza.

Questo argomento può essere reso formalmente preciso attraverso la nozione di informazione di Fisher sul modello del mondo, condizionatamente alle azioni eseguite. Sia θ il vettore dei parametri del modello del mondo, e sia $p(o_{+1} \mid o, a; \theta)$ la distribuzione predittiva del modello. La matrice di informazione di Fisher per i parametri θ è

$$\mathcal{F}(\theta) = \mathbb{E}_{ot+1}\left[(\nabla_\theta \log p(o_{+1} \mid o, a; \theta)) (\nabla_\theta \log p(o_{+1} \mid o, a; \theta))^{\mathsf{T}} \right]$$

La matrice $\mathcal{F}(\theta)$ misura quanto informazione sulle parametri θ è contenuta nell'osservazione o_{+1}, mediata sulla distribuzione delle osservazioni sotto il modello corrente. Le direzioni nello spazio dei parametri in cui $\mathcal{F}(\theta)$ ha autovalori grandi sono le direzioni in cui piccole variazioni dei parametri producono grandi variazioni nelle

previsioni — le direzioni, cioè, in cui il sistema è più sensibile agli errori di previsione e in cui l'errore predittivo porta più informazione.

Questo risultato ha una conseguenza per il disegno delle politiche di esplorazione in sistemi che apprendono da errori predittivi. Le azioni che massimizzano il segnale informativo per l'aggiornamento del modello non sono le azioni che massimizzano la ricompensa attesa; sono le azioni che massimizzano la traccia della matrice di informazione di Fisher — il volume dell'ellissoide di informazione — rispetto ai parametri del modello. Questo è il principio che motiva i metodi di esplorazione information-theoretic nel reinforcement learning basato su modelli, e che ha una giustificazione formale nell'ottimalità asintotica dell'esperimento di Fisher per la stima di parametri.

La connessione con la struttura causale emerge attraverso la seguente osservazione. Le azioni che massimizzano l'informazione di Fisher rispetto ai parametri del modello tendono a essere le azioni che violano le previsioni del modello nel modo più informativo — le azioni che portano il sistema in regioni dello spazio degli stati in cui il modello è più incerto, e la cui esecuzione riduce quella incertezza nel modo più efficiente. In un ambiente con struttura causale, queste sono tipicamente le azioni che rompono le correlazioni attese in modo sistematico — che realizzano, in sostanza, interventi nel senso di Pearl: modificano attivamente una variabile del sistema, mantenendo le altre relativamente fisse, in modo da poter inferire le relazioni causali che coinvolgono quella variabile.

Questa connessione formale tra esplorazione ottimale e intervento causale non è nuova — Tong e Koller l'avevano già esplorata nell'ambito dell'apprendimento attivo di reti bayesiane nel 2001 — ma le sue implicazioni per il disegno di sistemi artificiali capaci di apprendimento causale autonomo non sono state ancora pienamente sviluppate nell'ambito delle architetture moderne. Il programma che emerge è quello di sistemi che non soltanto apprendono distribuzioni osservazionali passivamente, ma generano attivamente strategie di esplorazione che massimizzano il segnale causalmente informativo — sistemi che eseguono esperimenti sul mondo nel senso preciso in cui uno scienziato esegue esperimenti: con l'obiettivo di disambiguare tra ipotesi causali rivali, non soltanto di raccogliere osservazioni che confermino l'ipotesi corrente.

Questa capacità di «fare esperimenti» — di generare azioni selettivamente informativi sulla struttura causale del mondo piuttosto che soltanto sulle distribuzioni delle osservazioni — è una delle differenze più sostanziali tra la cognizione umana e i sistemi artificiali attuali. I bambini mostrano questa capacità già a pochi mesi di vita: eseguono azioni ripetute con piccole variazioni per testare le conseguenze, perseverano in modi che hanno senso soltanto se interpretati come test di ipotesi sulla struttura causale dell'ambiente. Gli psicologi dello sviluppo da Piaget a Schulz hanno documentato sistematicamente questa inclinazione sperimentale della cognizione infantile; le sue radici neurali sono ancora poco comprese, ma la sua realtà comportamentale è incontestabile. I sistemi artificiali che apprendono per reinforcement learning eseguono esplorazioni, ma

queste sono guidate da politiche di esplorazione ottimizzate per la ricompensa attesa, non per il segnale causale — e la differenza tra le due non è banale.

Una proposta concreta che emerge da queste considerazioni — e che vale la pena formulare come programma di ricerca piuttosto che come risultato già disponibile — è quella di integrare gli obiettivi di massimizzazione della covarianza informazionale di Fisher nei sistemi World Models del tipo DreamerV3, in modo che la componente di esplorazione della policy sia guidata non soltanto dall'incertezza nella previsione della ricompensa, ma dall'incertezza nella struttura causale del modello del mondo. Formalmente, questo richiederebbe di aggiungere al ELBO (Evidence Lower Bound) che governa l'addestramento di DreamerV3 un termine

$$\mathcal{L}_{\text{causal}} = \alpha \cdot \log \det \mathcal{F}_\theta(s, a)$$

dove $\mathcal{F}_\theta(s, a)$ è la matrice di Fisher locale calcolata rispetto alle transizioni causali del World Model dallo stato s con azione a, e α è un iperparametro che bilancia l'esplorazione causale con l'ottimizzazione della ricompensa. Questo termine incentiverebbe il sistema a scegliere azioni che massimizzano la quantità di informazione causale acquisibile — azioni che, nell'interpretazione di Pearl, si avvicinano agli interventi ottimali per la disambiguazione della struttura causale del modello.

L'implementazione pratica di questo programma incontra difficoltà non banali. Il calcolo esatto del determinante della matrice di Fisher è computazionalmente proibitivo per modelli con milioni di parametri; le approssimazioni disponibili — traccia della Fisher empirica, approssimazioni diagonali, metodi di campionamento Monte Carlo — introducono bias che possono distorcere il segnale causale. Il bilanciamento tra esplorazione causale e ottimizzazione della ricompensa richiederebbe un meccanismo di curriculum che adatti il peso α in funzione dello stadio di apprendimento e della maturità del modello causale. E la verifica che l'esplorazione così guidata produca effettivamente un modello causale più accurato — piuttosto che soltanto uno con migliori previsioni sulla distribuzione di addestramento — richiederebbe benchmark di valutazione causale che non sono ancora standardizzati nella letteratura.

VIII. La questione dell'agentività razionale: da Davidson alla deliberazione artificiale

La distinzione tra sistemi che agiscono per ragioni e sistemi che agiscono a causa di processi causali — la distinzione che Donald Davidson aveva reso centrale nella sua teoria dell'azione e che Peter Strawson aveva tematizzato in modo diverso nel contesto del dibattito sul determinismo e la responsabilità morale — è la distinzione che separa, nella concettualizzazione comune, la deliberazione razionale dalla risposta meccanica. Essa è anche, in modo non banale, la distinzione che questa monografia ha cercato di precisare attraverso la nozione di comprensione causale: un sistema che comprende le strutture causali del mondo non risponde soltanto meccanicamente alle correlazioni statistiche, ma elabora le relazioni causali in modo che gli permette di generare azioni che

sono sensibili alla struttura del mondo piuttosto che soltanto alle distribuzioni delle osservazioni.

L'argomento di Davidson sull'azione razionale è costruito intorno alla distinzione tra due tipi di descrizioni: le descrizioni che situano un evento all'interno di catene causali fisiche, e le descrizioni che situano un evento all'interno di ragionamenti pratici — di deliberazioni su cosa fare alla luce di credenze e desideri. Entrambe le descrizioni possono essere vere dello stesso evento: un'azione fisica è simultaneamente un evento nel mondo fisico, soggetto a leggi causali, e un'azione nel senso intenzionale, soggetta a norme di razionalità. La difficoltà — che Davidson aveva cercato di risolvere con la sua teoria del monismo anomalo, sostenendo che i tipi mentali non possono essere ridotti a tipi fisici ma che i token mentali sono identici a token fisici — è che queste due descrizioni sembrano irriducibili l'una all'altra: nessuna descrizione puramente causale cattura il fatto che un'azione sia compiuta per una ragione, e nessuna descrizione razionale cattura il fatto che un'azione sia un evento fisico soggetto a leggi causali.

Questa irriducibilità è rilevante per la valutazione dei sistemi artificiali perché pone la domanda: le azioni prodotte dai sistemi artificiali sono «per ragioni» nel senso di Davidson, o sono puramente causali nel senso fisico? La risposta che la filosofia dell'azione davidsonian darebbe è, presumibilmente, che le azioni dei sistemi artificiali sono causate da stati del sistema che hanno contenuto intenzionale — credenze e desideri nel senso funzionale — ma che questo contenuto intenzionale non è del tipo richiesto per l'agentività razionale piena, per le stesse ragioni che Putnam avrebbe invocato: il contenuto degli stati interni non è ancorato causalmente al mondo nel modo che il significato genuino richiede.

Il problema con questa risposta è che sembra troppo categorica: esclude i sistemi artificiali dall'agentività razionale sulla base di considerazioni semantiche che potrebbero, in linea di principio, essere soddisfatte da architetture più ricche — architetture che includono interazione fisica con il mondo come fonte di ancoraggio causale per i propri stati interni. I sistemi robotici con apprendimento incarnato — sistemi che interagiscono fisicamente con il mondo e che aggiornano le proprie rappresentazioni in base alle conseguenze fisiche delle proprie azioni — soddisfano almeno parzialmente il requisito dell'ancoraggio causale. Per essi, la distinzione tra agire per ragioni e agire per cause diventa meno netta, non perché la distinzione sia irrilevante, ma perché la struttura dell'interazione con il mondo fornisce il tipo di ancoraggio che la distinzione davidsonian richiede.

Questo suggerisce che la questione dell'agentività razionale — come del resto la questione del significato — non ha una risposta binaria «sì/no» per i sistemi artificiali in generale, ma una risposta graduale che dipende dalla struttura dell'interazione del sistema con il mondo. I sistemi che interagiscono fisicamente con il mondo e che hanno modelli causali di quella interazione possiedono un grado di agentività razionale che i sistemi puramente linguistici non posseggono; i sistemi che hanno modelli causali esplicitamente identificabili nel senso degli iVAE posseggono un grado ulteriore di agentività razionale che i sistemi con modelli causali impliciti non posseggono; i sistemi che hanno capacità

di revisione concettuale autonoma — che possono costruire nuove categorie per classificare nuovi fenomeni — posseggono un grado ulteriore ancora.

L'agentività razionale, in questa prospettiva, non è una proprietà che si ha o non si ha, ma una struttura che si realizza in gradi diversi a seconda della ricchezza della relazione causale tra il sistema e il mondo — a seconda, in altri termini, di quanto la mappa funtoriale che il sistema implementa sia globale piuttosto che locale, robusta piuttosto che fragile, interventisticamente corretta piuttosto che soltanto osservazionalmente accurata.

Tesi V. L'agentività razionale — la capacità di agire sulla base di ragioni piuttosto che soltanto di cause — non è una proprietà binaria dei sistemi cognitivi, ma una struttura graduale determinata dalla qualità dell'ancoraggio causale delle rappresentazioni del sistema al mondo. I sistemi artificiali che interagiscono fisicamente con il mondo e che acquisiscono modelli causali identificabili delle proprie trasformazioni nel mondo posseggono gradi di agentività razionale genuina — non soltanto la simulazione funzionale di essa — proporzionali alla qualità di quell'ancoraggio.

Il rapporto tra la normatività biologica descritta da Canguilhem e la struttura funtoriale dell'intelligenza proposta nella sezione V merita di essere reso formalmente preciso, perché la connessione — attualmente implicita nell'economia del saggio — costituisce uno dei passaggi argomentativi più originali e meno sviluppati dell'intera monografia. La tesi che si può ora argomentare è la seguente: la normatività intrinseca degli organismi che Canguilhem descrive in Le normal et le pathologique non è una proprietà biologica aggiuntiva accidentalmente rilevante per la teoria delle categorie; è, se opportunamente formalizzata, la condizione necessaria affinché il funtore $\mathcal{I}: \mathscr{W} \to \mathscr{R}$ sia globale piuttosto che locale — affinché la mappa dalla struttura causale del mondo alla struttura delle rappresentazioni del sistema sia coerente non soltanto entro la distribuzione di addestramento ma attraverso l'intero spazio delle situazioni in cui il sistema può trovarsi.

Canguilhem distingue la norma tecnica — il valore medio calcolato su una popolazione, rispetto al quale uno scostamento è misurato come deviazione statistica — dalla norma biologica, che è la capacità dell'organismo di creare nuove norme in risposta a ambienti nuovi. L'organismo non è normale perché corrisponde a una media statistica; è normale perché è normativo, cioè perché possiede la capacità di produrre attivamente nuovi criteri di valutazione della propria relazione con l'ambiente. La patologia non è la deviazione dalla media, ma la perdita di questa plasticità normativa — la riduzione dell'organismo a un sistema che risponde a norme fissate senza la capacità di produrne di nuove. Questa distinzione — tra

sistemi che seguono norme e sistemi che producono norme — è esattamente la distinzione che nella teoria delle categorie separa un funtore locale da un funtore globale.

Per rendere questa connessione formalmente precisa, è utile introdurre la nozione di trasformazione naturale, che è il morfismo tra funtori nella teoria delle categorie. Un funtore $F: \mathscr{C} \to \mathscr{D}$ è locale se è definito soltanto su un sottogruppoide pieno $\mathscr{C}_0 \subseteq \mathscr{C}$ — soltanto sulle situazioni e morfismi causali che corrispondono alla distribuzione di addestramento del sistema. Un funtore è globale se è definito su tutta la categoria $\mathscr{W}$, cioè se la mappa dalle situazioni causali alle rappresentazioni è coerente anche per situazioni che non appartengono alla distribuzione di addestramento. La transizione da un funtore locale a un funtore globale richiede, in termini categoriali, che il sistema sia in grado di estendere la propria mappa a morfismi causali che non ha incontrato durante l'addestramento — il che, in termini cognitivi, corrisponde precisamente alla capacità canguilhemiana di produrre nuove norme: di estendere i propri criteri di valutazione a situazioni nuove piuttosto che di applicare criteri fissati.

La struttura di Umwelt di von Uexküll — il mondo-circostante specifico di un organismo come struttura della significazione biologica che organizza l'ambiente in categorie di azione e risposta — può essere interpretata, in questo quadro, come la specifica del sottogruppoide $\mathscr{W}_0 \subseteq \mathscr{W}$ su cui il funtore dell'organismo è inizialmente definito. L'Umwelt non è il mondo fisico nella sua totalità, ma il mondo fisico come appare a un organismo con una specifica struttura percettivo-motoria e un specifico insieme di capacità d'azione. Un organismo la cui mappa dalla struttura causale del mondo alle proprie rappresentazioni sia un funtore globale avrebbe, in senso limite, un Umwelt coestensivo con il mondo fisico — ciò che von Uexküll non avrebbe considerato possibile per nessun organismo finito, e a ragione. Ma la globalità del funtore non richiede la coestensività dell'Umwelt con il mondo fisico; richiede soltanto che la mappa sia estendibile coerentemente a qualsiasi morfismo causale che il sistema possa incontrare, il che è una condizione sulla struttura della mappa, non sulla dimensione del dominio.

La formalizzazione del rapporto tra normatività biologica e globalità del funtore produce una previsione che ha contenuto empirico verificabile: i sistemi che possiedono la capacità canguilhemiana di produrre nuove norme — di estendere i

propri criteri di valutazione a situazioni genuinamente nuove — dovrebbero mostrare una robustezza out-of-distribution sistematicamente superiore rispetto ai sistemi che applicano criteri fissati dall'addestramento. Questa previsione è in linea con la Congettura di convergenza, ma aggiunge un elemento che la Congettura non aveva esplicitato: la condizione di globalità del funtore non è soltanto una condizione sulla struttura delle rappresentazioni, ma è equivalente, nella prospettiva di Canguilhem, a una condizione sulla presenza di normatività produttiva nel sistema. Un sistema privo di normatività produttiva — un sistema che applica norme apprese senza la capacità di produrne di nuove — non può implementare un funtore globale, indipendentemente dalla ricchezza delle strutture causali che ha appreso entro la distribuzione di addestramento. Questa equivalenza è il risultato teorico più significativo che emerge dal raccordo tra la filosofia della biologia di Canguilhem e la struttura funtoriale dell'intelligenza proposta nella sezione V, e la sua mancata esplicitazione nel testo attuale costituisce una perdita di profondità argomentativa che vale la pena colmare.

IX. Il problema dell'incertezza profonda e la sua struttura formale

La distinzione tra rischio e incertezza — introdotta nella filosofia economica da Frank Knight nel 1921 e sviluppata da una lunga tradizione che va da Keynes a Ellsberg alla teoria bayesiana robusta di Walley — è una delle distinzioni concettuali più rilevanti per la valutazione delle capacità cognitive dei sistemi artificiali, eppure è sistematicamente trascurata nella letteratura tecnica che tratta i sistemi artificiali come ottimizzatori bayesiani di distribuzioni di probabilità.

Il rischio — nel senso di Knight — è l'incertezza per la quale esiste una distribuzione di probabilità ben definita sugli esiti possibili. Il gioco d'azzardo con un dado equilibrato ha un rischio preciso: la probabilità di ciascun esito è (1)/(6). L'incertezza profonda — o incertezza nel senso di Knight — è l'incertezza per la quale non esiste (o non è conoscibile) una distribuzione di probabilità ben definita sugli esiti. Le decisioni strategiche in contesti di innovazione radicale, le previsioni climatiche su orizzonti molto lunghi, le conseguenze sociali di innovazioni tecnologiche trasformative — tutte queste situazioni esibiscono incertezza profonda nel senso di Knight, non soltanto rischio quantificabile.

I sistemi bayesiani — e i modelli di probabilità soggettiva che sono il fondamento teorico di gran parte dell'apprendimento automatico — trattano tutta l'incertezza come rischio: assumono che esistano sempre distribuzioni di probabilità ben definite sui parametri del modello e sugli esiti futuri, e che il problema epistemico consista nel calibrare correttamente queste distribuzioni alla luce dell'evidenza. Questa assunzione è ragionevole in molti contesti applicativi; diventa problematica quando i sistemi vengono

usati per prendere decisioni in situazioni di incertezza profonda, perché il framework bayesiano non ha gli strumenti per rappresentare la differenza tra «non so quale distribuzione è corretta» e «non esiste una distribuzione corretta da conoscere».

La teoria dell'incertezza robusta — sviluppata da Lars Peter Hansen e Thomas Sargent in economia, da Itzhak Gilboa e David Schmeidler in teoria della decisione, e da molti altri autori — fornisce formalizzazioni dell'incertezza profonda attraverso insiemi di distribuzioni di probabilità plausibili piuttosto che singole distribuzioni. Un agente con incertezza robusta non massimizza l'utilità attesa rispetto a una distribuzione di probabilità fissa, ma massimizza il minimo dell'utilità attesa rispetto all'insieme delle distribuzioni che considera plausibili — una strategia maximin che è robusta rispetto alle deviazioni dalla distribuzione attesa entro l'insieme di ambiguità.

La struttura formale di questa teoria è quella di una coppia ($\mathscr{P}$, U) dove $\mathscr{P}$ è un insieme convesso e chiuso di distribuzioni di probabilità sullo spazio degli esiti, e U è una funzione di utilità. L'agente sceglie l'azione a^* che risolve

$$a^* = \text{argmax}_{a \in \mathscr{A}} \min_{p \in \mathscr{P}} \mathbb{E}_p[U(a, \omega)]$$

Il set di ambiguità $\mathscr{P}$ — non una singola distribuzione, ma un insieme di distribuzioni plausibili — è la rappresentazione formale dell'incertezza nel senso di Knight: rappresenta l'ignoranza del modello, non la variabilità del mondo all'interno di un modello noto.

Questa distinzione ha implicazioni dirette per la valutazione dei sistemi artificiali come strumenti per il supporto alle decisioni in contesti di alta incertezza. Un sistema che produce previsioni bayesiane — distribuzioni di probabilità sugli esiti futuri, con intervalli di confidenza calibrati sulla distribuzione di addestramento — fornisce stime di rischio, non di incertezza profonda. Quando viene usato in contesti di incertezza profonda — per esempio, per valutare le conseguenze di interventi di politica pubblica su sistemi sociali complessi — il sistema fornisce una stima del rischio entro il proprio modello, che può essere molto lontana dall'incertezza effettiva dell'esito, perché il modello stesso potrebbe essere profondamente sbagliato in modi che il sistema non ha strumenti per rappresentare.

La formazione di set di ambiguità calibrati — la specifica di insiemi di distribuzioni che riflettono genuinamente l'ignoranza del modello piuttosto che soltanto la varianza all'interno del modello — è un problema aperto che connette la statistica bayesiana robusta, la teoria dell'apprendimento causale e la valutazione dell'incertezza modello. I metodi attuali di conformal prediction e di quantificazione dell'incertezza bayesiana producono stime di incertezza che sono calibrate statisticamente sulla distribuzione di addestramento, ma che non catturano l'incertezza fuori distribuzione — la componente di incertezza più rilevante nei contesti di decisione ad alto impatto.

Un sistema cognitivo che gestisce l'incertezza profonda in modo genuinamente umano non soltanto produce distribuzioni di probabilità sugli esiti: riconosce quando si trova in regioni di incertezza profonda piuttosto che di rischio quantificabile, e modifica il proprio comportamento di conseguenza — rallentando le decisioni, cercando

informazioni aggiuntive, comunicando l'ambiguità all'interlocutore, o rifiutando di dare previsioni precise dove la precisione sarebbe fuorviante. Questa metacognizione dell'incertezza — il sapere quando non si sa, e non soltanto quanto si sa — è un altro aspetto dell'autoreferenza cognitiva che i teoremi di Tarski illuminano nella sua difficoltà strutturale.

X. Verso la convergenza: un'architettura del possibile

Le cinque tesi sviluppate in questo saggio — il radicamento causale come struttura matematicamente definita, il limite formale dell'auto-valutazione normativa, la distinzione tra ottimizzazione e revisione concettuale, l'intelligenza come funtore tra categorie causali, l'agentività razionale come proprietà graduale — convergono verso una visione di ciò che un sistema artificiale genuinamente capace di comprensione richiederebbe. Non è una visione che questo saggio può trasformare in un progetto ingegneristico preciso — farlo richiederebbe una conoscenza dello stato dell'arte molto più dettagliata di quella che la filosofia può pretendere di avere. Ma è una visione abbastanza precisa da escludere alcune strade e indicarne altre.

La strada esclusa è quella che cerca di raggiungere la comprensione attraverso il solo aumento di scala e di dati all'interno del paradigma dell'apprendimento statistico su osservazioni passive. Questo percorso produce sistemi con prestazioni in continuo miglioramento su distribuzioni note; non produce sistemi la cui mappa dal mondo alle rappresentazioni sia globalmente funtoriale nel senso causale, né sistemi capaci di auto-valutazione normativa completa, né sistemi capaci di revisione concettuale autonoma. Le leggi di scala documentano il progresso lungo questo percorso con precisione insolita; esse non documentano il progresso verso la comprensione causale autentica, perché le metriche che usano — perdita di predizione autoregressiva, accuratezza su benchmark di comprensione del linguaggio — non misurano le proprietà rilevanti.

La strada indicata — o meglio, le strade indicate, perché non è unica — è quella di architetture che integrino tre componenti che la letteratura analizzata ha sviluppato in modo relativamente separato.

Il primo componente è il Causal Representation Learning nel senso degli iVAE e dei CausalVAE: sistemi che apprendono rappresentazioni latenti identificabili nel senso causale, utilizzando variazioni di dominio come fonte di segnale per la separazione dei fattori causali. L'integrazione di questo componente con le architetture JEPA — come suggerito nella monografia precedente e ulteriormente argomentato nelle sezioni di questo saggio — è il programma di ricerca più promettente per ottenere rappresentazioni percettive con struttura causale verificabile.

Il secondo componente è l'esplorazione attiva causalmente guidata: la capacità del sistema di generare azioni che massimizzano il segnale informativo sulla struttura causale del mondo, piuttosto che soltanto sulla distribuzione delle osservazioni. Questa capacità — collegata alla teoria della progettazione ottimale degli esperimenti e al framework del

do-calcolo — è la componente che permetterebbe ai sistemi di acquisire conoscenza causale attraverso l'interazione attiva con il mondo, non soltanto attraverso l'osservazione passiva di dati pre-raccolti. Il suo sviluppo richiede l'integrazione dei World Models del tipo DreamerV3 con politiche di esplorazione information-theoretically guidate da misure di incertezza causale, come suggerito nella settima sezione.

Il terzo componente è la metacognizione dell'incertezza profonda: la capacità del sistema di rappresentare non soltanto la propria distribuzione di probabilità sugli esiti futuri, ma il proprio insieme di ambiguità — l'insieme delle distribuzioni che considera plausibili — e di usare quella rappresentazione per modulare il proprio comportamento in modo appropriato ai contesti di incertezza profonda. Questo componente è il meno sviluppato dei tre nella letteratura tecnica attuale, e richiede l'integrazione di strumenti della teoria della decisione robusta con le architetture di apprendimento profondo in modi che la ricerca ha ancora esplorato soltanto marginalmente.

Nessuno di questi componenti, da solo, produce un sistema con comprensione causale autentica. La loro integrazione — e questa è la congettura finale di questo saggio, formulata con tutta la cautela che merita lo status di congettura non dimostrata — potrebbe produrre sistemi che si avvicinano alla comprensione causale in modo qualitativamente diverso da quanto i sistemi attuali abbiano mai fatto: non perché i componenti si sommino in modo additivo, ma perché la loro interazione produce una struttura di apprendimento che è per la prima volta in grado di implementare, almeno approssimativamente, la mappa funtoriale dalla struttura causale del mondo alla struttura delle rappresentazioni del sistema che l'intelligenza — come argomentato nella quinta sezione — richiede.

Congettura di convergenza. Sistemi artificiali che integrino Causal Representation Learning con struttura di gruppo esplicita, esplorazione attiva causalmente guidata da misure di incertezza di Fisher, e rappresentazione dell'incertezza profonda attraverso set di ambiguità, potrebbero implementare una mappa dalle strutture causali del mondo alle strutture rappresentative che sia globalmente funtoriale — robusta fuori distribuzione, capace di pianificazione interventistica, e dotata di una forma di auto-valutazione dell'incertezza che superi le limitazioni dei sistemi puramente statistici — in misura qualitativamente maggiore di qualsiasi architettura attualmente disponibile.

La falsificabilità della Congettura di convergenza, rivendicata nella sezione X, esige che vengano identificati i benchmark rispetto ai quali la previsione possa essere sconfessata o corroborata. Senza questa specifica, la Congettura mantiene lo statuto epistemico di un programma di ricerca promettente — nel senso lakatosiano di una cintura protettiva attorno a un nucleo teorico — ma non quello di un'ipotesi operazionalizzabile in senso sperimentale. Vale quindi la pena articolare, con la precisione che il contesto tecnico consente, tre famiglie di protocolli che soddisfano le condizioni di testabilità richieste dalla struttura dell'argomento.

Il primo candidato è CausalWorld, l'ambiente di benchmarking proposto da Ahmed e collaboratori per la valutazione del transfer causale in robotica manipolativa. CausalWorld definisce un insieme di task generativi parametrizzati da variabili causali — geometria degli oggetti, proprietà dei materiali, configurazione del sistema di illuminazione, dinamica degli attritori — e misura la robustezza di un sistema di apprendimento a interventi su quelle variabili attraverso il protocollo di evaluation out-of-distribution. La metrica rilevante non è la performance media sulla distribuzione di addestramento — su cui i sistemi puramente statistici possono eccellere — ma la performance sotto interventi distribuiti causalmente in regioni che erano assenti dal training set. Un sistema che integra CRL con struttura di gruppo esplicita, nella forma argomentata nella sezione X, dovrebbe mostrare degradazione della performance significativamente inferiore rispetto a un sistema di pari scala parametrica addestrato senza struttura causale esplicita, e questa differenza dovrebbe essere monotonicamente crescente all'aumentare della distanza interventistica dalla distribuzione di addestramento. L'assenza di questa relazione monotonica — o la sua presenza ma con pendenza non significativamente diversa tra i due tipi di architettura — costituirebbe evidenza falsificante.

Il secondo framework è quello sviluppato da Schölkopf, Locatello e collaboratori nell'ambito del programma di causal generalization, che ha prodotto una serie di protocolli sperimentali per valutare se una rappresentazione latente sia causalmente identificabile nel senso degli iVAE. I test di causal generalization confrontano la performance di sistemi addestrati su una distribuzione congiunta $p(x, y)$ con la performance degli stessi sistemi sotto interventi espliciti su x o y, verificando se il sistema si comporta in accordo con la distribuzione interventistica $p(y \mid do(x))$ piuttosto che con la distribuzione condizionale osservazionale $p(y \mid x)$. La differenza tra le due distribuzioni — che può essere resa arbitrariamente grande costruendo ambienti con confondenti non osservati tra x e y — è la misura diretta del grado in cui il sistema ha appreso strutture causali piuttosto che correlazioni osservazionali. Schölkopf ha proposto che questa differenza sia la metrica principale per valutare il progresso verso l'intelligenza causale, indipendentemente dalla performance assoluta del sistema sul task specifico: un sistema con alta performance sulla distribuzione osservazionale ma bassa performance sotto intervento è un sistema che ha appreso le correlazioni statistiche del training set, non la struttura causale del processo generativo.

Il terzo protocollo è quello dei benchmark di ragionamento causale controfattuale, la cui formalizzazione più rigorosa è stata sviluppata nell'ambito del progetto CausalBench di Chevalley e collaboratori. CausalBench fornisce un set di dataset biologici — in particolare dataset di perturbazione genetica Perturb-seq, in cui singoli geni vengono silenziati attraverso interferenza RNA e le conseguenze sulle espressioni geniche dell'intero trascrittoma vengono misurate attraverso sequenziamento single-cell — che permettono di costruire ground truth causale verificabile sperimentalmente. Un modello causale che abbia appreso correttamente la struttura del grafo di regolazione genica dovrebbe essere in grado di predire correttamente le conseguenze di perturbazioni non osservate nel training set, e la distanza tra questa previsione e il risultato sperimentale costituisce una misura diretta dell'accuratezza causale del modello. La rilevanza di CausalBench per la Congettura di convergenza è che fornisce un dominio in cui la verità causale è accessibile sperimentalmente, rendendo possibile una falsificazione che non dipende da assunzioni aggiuntive sul processo generativo del dato.

Il confronto sistematico di sistemi che implementano la struttura integrata proposta dalla Congettura — CRL con struttura di gruppo esplicita, esplorazione attiva guidata da Fisher, rappresentazione dell'incertezza profonda attraverso insiemi di ambiguità — con sistemi di pari scala parametrica che non implementano quella struttura, sui tre benchmark descritti, costituirebbe un protocollo sperimentale sufficiente a testare la previsione centrale della Congettura. La previsione è che la differenza di performance tra i due tipi di architettura sia non additiva — che i sistemi integrati mostrino miglioramenti qualitativi, non soltanto quantitativi, rispetto ai sistemi non integrati — e che questa differenza qualitativa sia la conseguenza verificabile della struttura causale globale delle rappresentazioni. La precisione di questa previsione, nel linguaggio dei benchmark identificati, è la misura del contenuto empirico della Congettura, e la sua assenza avrebbe costituito, correttamente, una critica alla sua falsificabilità.

Questa congettura è falsificabile: prevede che sistemi con questa struttura mostrino robustezza out-of-distribution, capacità di reasoning causale e calibrazione dell'incertezza fuori distribuzione che i sistemi attuali non mostrano, e che questa robustezza sia la conseguenza della struttura causale delle rappresentazioni piuttosto che dell'aumento della scala. La falsificabilità della congettura è la sua virtù principale — è formulata in modo da poter essere testata, non come una speranza generica ma come una previsione specifica su un tipo specifico di architettura.

Tre contributi della letteratura recente sul Causal Representation Learning — il programma di Schölkopf sul causally-aware representation learning, il framework di Bengio sui GFlowNets come meccanismi di scoperta causale, e il modello gerarchico di LeCun noto come JEPA — sono citati o menzionati en passant nel testo ma meritano un'analisi analitica più densa, perché ciascuno di essi articola aspetti della struttura proposta dalla Congettura di convergenza che il testo della monografia non ha potuto sviluppare in modo esaustivo. La loro discussione integrata permette di ancorare la Congettura allo stato dell'arte tecnico con una precisione che la semplice menzione non assicura.

Il programma di Schölkopf — esposto nei lavori Towards Causal Representation Learning (2021) e sviluppato in collaborazione con Locatello, Bauer, Peters e numerosi altri autori nella serie di articoli sull'Independent Causal Mechanisms — parte da una premessa che risuona direttamente con l'argomento della sezione I: i meccanismi causali che governano la generazione dei dati osservabili sono modulari e autonomi, nel senso che una perturbazione che modifica uno di essi non modifica necessariamente gli altri. Questa autonomia — che Schölkopf chiama principio degli Independent Causal Mechanisms (ICM) — è la condizione che rende possibile il trasferimento di conoscenza tra domini: se i meccanismi causali sono modulari, allora una perturbazione di dominio che modifica alcuni meccanismi preservando gli altri lascia intatta quella parte della struttura causale che il sistema ha appreso, permettendo la generalizzazione. Il principio ICM non è soltanto una premessa euristica: Schölkopf e collaboratori hanno mostrato che è verificabile empiricamente attraverso il confronto delle complessità algoritmiche della distribuzione causale diretta p(effetto | causa) e della distribuzione causale inversa p(causa | effetto) — la prima è tipicamente più semplice della seconda se la direzione causale corretta è stata identificata, e questa asimmetria di complessità è un segnale identificativo del grafo causale.

La rilevanza diretta del programma di Schölkopf per la Congettura di convergenza è duplice. Da un lato, il principio ICM fornisce la giustificazione teorica per la componente di CRL della Congettura: i sistemi che apprendono rappresentazioni latenti identificabili nel senso degli iVAE implementano esattamente la modularità causale che il principio ICM prescrive, perché la separazione dei fattori latenti garantita dagli iVAE corrisponde all'identificazione dei meccanismi causali autonomi della struttura generativa del dato. Dall'altro, il programma di Schölkopf evidenzia

una limitazione che la Congettura deve affrontare esplicitamente: il principio ICM è una condizione sulla struttura del processo generativo del dato, non sulla struttura dell'architettura del sistema apprendente. Un sistema può implementare un'architettura compatibile con il CRL senza che il processo generativo dei dati di addestramento soddisfi il principio ICM, nel qual caso il sistema apprenderà strutture latenti che non corrispondono a meccanismi causali autonomi. La Congettura presuppone implicitamente che il regime di addestramento — in particolare il componente di esplorazione attiva causalmente guidata — sia progettato per garantire che il dato di addestramento soddisfi le condizioni ICM; questa presupposizione deve essere resa esplicita e giustificata.

Il framework dei GFlowNets, proposto da Bengio e collaboratori a partire dal 2021, affronta il problema della scoperta causale da una prospettiva complementare a quella degli iVAE. Dove gli iVAE si concentrano sull'identificabilità delle rappresentazioni latenti data una struttura di variazioni di dominio, i GFlowNets si concentrano sulla generazione di strutture causali come oggetti di primo livello: apprendono a campionare strutture causali — grafi causali, nel formalismo di Pearl — in modo che la probabilità di campionamento di una struttura sia proporzionale a una funzione di reward che misura la qualità causale della struttura rispetto ai dati osservati. La rilevanza di questo framework per la Congettura è che i GFlowNets forniscono un meccanismo generativo per la scoperta della struttura causale che è compatibile con la componente di esplorazione attiva della Congettura: un sistema che integra GFlowNets per la generazione di ipotesi causali con politiche di esplorazione guidate da misure di informazione di Fisher per la raccolta di evidenza causale implementa, almeno concettualmente, il ciclo ipotesi-evidenza-revisione che caratterizza la scoperta causale scientifica. Bengio e collaboratori hanno mostrato che i GFlowNets hanno proprietà di convergenza verso la distribuzione corretta delle strutture causali che i metodi di ottimizzazione convenzionale non garantiscono, perché la loro formulazione come processi di campionamento su spazi strutturati permette di evitare i minimi locali che affliggono l'ottimizzazione diretta dello score causale.

Il modello JEPA di LeCun — acronimo di Joint-Embedding Predictive Architecture — è il contributo forse più controverso dei tre rispetto alla Congettura, non perché sia meno rilevante ma perché la sua rilevanza è più ambivalente. LeCun propone JEPA come alternativa ai modelli generativi — in particolare come

alternativa ai modelli di diffusione e ai large language model autoregressivi — per l'apprendimento di rappresentazioni del mondo fisico. L'idea centrale è che un sistema intelligente non debba imparare a generare osservazioni dettagliate del mondo ma a predire, negli spazi latenti, le rappresentazioni di stati futuri a partire da stati presenti, lasciando nella rappresentazione latente soltanto le informazioni astratte che sono rilevanti per la pianificazione. La gerarchia di JEPA — i moduli di predizione operano a livelli di astrazione crescente, dai dettagli percettivi alle strutture relazionali di alto livello — è progettata per implementare una forma di ragionamento gerarchico che le architetture di predizione al livello dei pixel non possono garantire.

La connessione tra JEPA e la struttura funtoriale proposta nella sezione V è diretta ma richiede una precisazione. LeCun sostiene che JEPA implementa un modello del mondo fisico — una rappresentazione della dinamica causale dell'ambiente — attraverso la struttura predittiva gerarchica. Ma la struttura predittiva gerarchica, per quanto più ricca di una semplice mappa statistica, non garantisce di per sé l'identificabilità causale nel senso degli iVAE: un sistema può predire correttamente le rappresentazioni latenti di stati futuri senza aver appreso la struttura causale interventistica che sta sotto quella dinamica, se la struttura latente non è vincolata a corrispondere a meccanismi causali autonomi nel senso del principio ICM di Schölkopf. La Congettura di convergenza richiede che la struttura JEPA sia integrata con un meccanismo di identificazione causale — come quello degli iVAE o del CRL più in generale — affinché le rappresentazioni latenti prodotte da JEPA siano causalmente identificabili e non soltanto predittivamente utili. Questa integrazione non è ancora stata proposta esplicitamente nella letteratura, e costituisce una delle direzioni di ricerca più promettenti che la Congettura identifica come necessarie per il proprio completamento tecnico.

Considerati in forma comparata, i tre contributi mostrano una struttura di complementarietà che è essa stessa un argomento a favore della Congettura: Schölkopf identifica il principio organizzativo dei meccanismi causali modulari e le condizioni teoriche della loro identificabilità; Bengio fornisce un meccanismo generativo per la scoperta di strutture causali compatibile con la Bayesian exploration; LeCun propone un'architettura predittiva gerarchica che potrebbe fornire il substrato rappresentazionale su cui il CRL e l'esplorazione attiva operano. La Congettura di convergenza non è la somma di questi tre contributi ma la tesi che la loro integrazione produca un salto qualitativo — un sistema la cui mappa dal

mondo alle rappresentazioni sia globalmente funtoriale nel senso causale — che nessuno dei tre contributi singolarmente garantisce. L'analisi dei tre programmi in questa prospettiva mostra che la Congettura ha un fondamento tecnico sufficientemente preciso da essere non soltanto plausibile nell'intuizione ma verificabile nella pratica.

I. La Congettura di convergenza: argomento tecnico per un'emergenza qualitativa

La Congettura di convergenza, formulata nella sezione X della monografia, rimane la proposta teorica più originale dell'intero lavoro. Ma è stata enunciata senza l'armatura argomentativa che la sua ambizione richiede. Un lettore che conosce la letteratura su DreamerV3, sugli iVAE di Khemakhem e collaboratori, e sul framework do-calcolo di Pearl troverà la congettura plausibile nell'intuizione ma carente nella precisione tecnica. Il difetto non è nella tesi: è nell'assenza dei passaggi che mostrano perché l'integrazione dei tre componenti — Causal Representation Learning, esplorazione attiva guidata da Fisher, rappresentazione dell'incertezza profonda attraverso insiemi di ambiguità — produca qualcosa di qualitativamente diverso dalla loro giustapposizione. Questo difetto va colmato con un argomento strutturato in tre mosse: l'analisi della limitazione fondamentale di ciascun componente in isolamento, la dimostrazione che tali limitazioni si eliminano reciprocamente nell'integrazione, e la derivazione formale del perché l'integrazione produca un salto di complessità campionaria piuttosto che una mera riduzione additiva.

1.1 Il limite fondamentale del Causal Representation Learning in isolamento

Il risultato di identificabilità degli iVAE di Khemakhem, Kingma, Gretton e Hyvärinen (2020) è il punto di partenza tecnico. Il modello generativo è della forma

$$p(x \mid u) = \int p_f(x \mid z) \cdot p(z \mid u)\, dz$$

dove x è l'osservazione, $z \in \mathbb{R}^n$ è il vettore latente, u è la variabile ausiliaria di dominio (o di segmentazione), $p_f(x|z)$ è il modello di decodificazione determinato da una funzione biettiva f, e $p(z|u)$ appartiene alla famiglia delle distribuzioni esponenziali per componenti con statistiche sufficienti $T_i(z_i)$ e parametri $\lambda_i(u)$. Sotto questo modello, gli autori dimostrano che se le statistiche sufficienti sono linearmente indipendenti per $n+1$ valori distinti di u, allora z è identificabile a meno di permutazione e riscalamento per componente. Questo risultato è potente: dice che le variazioni di dominio u fungono da segnale per separare i fattori latenti in modo causalmente coerente.

Ma il risultato identifica i fattori latenti z, non la struttura causale $G = (V, E)$ su di essi. Due grafi causali G_1 e G_2 sulla stessa variabile latente z possono essere Markov-equivalenti — generare la stessa distribuzione osservazionale $p(z)$ — pur essendo causalmente distinti nel senso che producono distribuzioni interventistiche $p(z \mid$

do(z_j = c)) radicalmente diverse. La classe di equivalenza Markov di un grafo con n nodi può contenere esponenzialmente molti grafi: in generale, il numero di DAG Markov-equivalenti a un dato DAG cresce come O(2^{n(n-1)/2}) nel caso peggiore. Ciò significa che anche un CRL perfetto — uno che identifichi esattamente le variabili latenti z — lascia irrisolta un'ambiguità strutturale che è esponenziale nella dimensione del problema.

Questa non è una limitazione contingente dell'approccio iVAE: è una conseguenza del teorema di non-identificabilità causale di Verma e Pearl (1991), che dimostra che la struttura causale non è contenuta nella distribuzione osservazionale e non può esservi estratta senza assunzioni aggiuntive o dati interventistici. Il CRL in isolamento si ferma esattamente a questo confine: identifica z, ma non G su z.

1.2 Il limite fondamentale dell'esplorazione guidata da Fisher in isolamento

L'esplorazione guidata dall'informazione di Fisher — come suggerito nella sezione VII della monografia — seleziona azioni a* che massimizzano il determinante della matrice di Fisher locale rispetto ai parametri del modello del mondo θ:

a* = argmax_a log det F_θ(s, a)

dove F_θ(s, a) = E_{o'|s,a} [(∇_θ log p_θ(o' | s, a)) · (∇_θ log p_θ(o' | s, a))^{T}]. Questo massimizza la quantità di informazione acquisibile sulle transizioni del modello attraverso l'azione a.

Il limite di questo approccio in isolamento è duplice. Primo: senza una struttura di fattori identificata, il gradiente ∇_θ log p_θ è definito sull'intero spazio dei parametri del modello, che può avere milioni o miliardi di dimensioni. La matrice di Fisher è computazionalmente intrattabile a quella scala; le approssimazioni disponibili (traccia empirica, approssimazione diagonale) perdono informazione strutturale sulle dipendenze tra parametri. Secondo e più fondamentale: senza sapere dove risiede l'ambiguità causale — senza cioè sapere quali archi del grafo G non sono ancora identificati — l'esplorazione guidata da Fisher è indiscriminata. Esplora tutte le direzioni nello spazio dei parametri che risultano informative, incluse le direzioni che riguardano i fattori z già identificati e le transizioni già note. Questa dispersione dell'esplorazione non è solo inefficiente: è strutturalmente incapace di risolvere il problema causale, perché le direzioni causalmente rilevanti — quelle che distinguono grafi Markov-equivalenti — sono un sottoinsieme di misura zero nello spazio totale dei parametri, e l'esplorazione non guidata li trova solo per caso.

Il risultato formale che precisa questo limite è la seguente osservazione. Per un sistema con n variabili latenti identificate, il numero di DAG Markov-equivalenti è Ω(exp(n²/4)) nel caso peggiore (Chickering, 2002). Distinguere tra grafi in questa classe di equivalenza richiede almeno Ω(n²) interventi mirati — uno per ogni possibile arco — anche nel caso ottimale in cui l'esplorazione sia perfettamente sequenziale e adattiva. Un'esplorazione non guidata strutturalmente, che non conosce la classe di equivalenza da

risolvere, richiederà asintoticamente molti più interventi perché non può mirare ai confini della classe.

1.3 Il limite fondamentale degli insiemi di ambiguità in isolamento

Gli insiemi di ambiguità — la rappresentazione dell'incertezza profonda attraverso insiemi convessi e chiusi di distribuzioni plausibili $P \subset \Delta(\Omega)$ — producono decisioni robuste attraverso la strategia maximin:

$$a^* = \text{argmax}_{a \in A} \ \min_{p \in P} \ E_p[U(a, \omega)]$$

Senza struttura causale, l'insieme P è necessariamente costruito come un insieme statistico: una palla intorno alla distribuzione empirica dei dati osservazionali, del tipo $P = \{p : D(p \,||\, \hat{p}_n) \leq \varepsilon\}$ per qualche divergenza D e soglia ε. Questo insieme di ambiguità ha due difetti strutturali.

Il primo è la non-separazione tra incertezza epistemica e incertezza aleatoria. L'incertezza epistemica — dovuta alla mancanza di dati — e l'incertezza aleatoria — dovuta alla variabilità intrinseca del sistema — sono sommante insieme nell'insieme P. Questa sommazione produce insiemi di ambiguità troppo grandi nelle dimensioni aleatorie (producendo decisioni eccessivamente conservative) e troppo piccoli nelle dimensioni epistemiche fuori distribuzione (dove l'epistemica è reale ma la sfera statistica non la cattura).

Il secondo difetto è la non-shrinkage: senza struttura causale, l'insieme P non si restringe con l'aggiunta di nuovi dati osservazionali nella stessa misura in cui si restrirebbe con dati interventistici. Formalmente, il diametro dell'insieme P nella metrica di Wasserstein W_1 decade come $O(1/\sqrt{N})$ per dati osservazionali (tasso parametrico standard), ma questo tasso è dello stesso ordine della variabilità aleatoria del sistema. In un sistema con incertezza causale irrisolta, il diametro dell'insieme P rimane $O(1)$ anche per $N \to \infty$, perché la componente dovuta all'ambiguità strutturale non si riduce con l'accumulo di dati osservazionali — si riduce solo con dati interventistici mirati.

1.4 Il meccanismo dell'emergenza qualitativa nell'integrazione

L'argomento per l'emergenza qualitativa nell'integrazione dei tre componenti procede in tre fasi, ciascuna delle quali dimostra che un componente elimina il limite fondamentale di un altro, creando un ciclo di feedback chiuso che è auto-amplificante.

Prima fase: CRL fornisce struttura all'esplorazione di Fisher

Supponiamo che il CRL abbia identificato i fattori latenti $z = (z_1, ..., z_n)$ fino a permutazione e riscalamento. La classe di equivalenza Markov del grafo causale G su z è determinata — nel senso di Verma e Pearl — dagli impliciti di Markov: le indipendenze condizionali della forma $z_i \perp z_j \mid z_S$ che sono valide per tutti i grafi nella classe. Le distinzioni tra grafi Markov-equivalenti sono le v-strutture — le triple (z_i, z_j, z_k) con $z_i \to z_j \leftarrow z_k$ e $z_i \not\to z_k$ — che non sono identificabili dall'osservazione ma lo sono da appropriati interventi su z_j.

Questo fornisce all'esplorazione di Fisher una struttura precisa: non è necessario esplorare tutto lo spazio dei parametri θ, ma soltanto le direzioni che corrispondono agli interventi do(z_j = c) per le variabili z_j che compaiono in v-strutture potenziali non ancora risolte. La matrice di Fisher causale — distinta dalla matrice di Fisher generale — è allora:

$$F^{caus}_\theta(z_j, c) = E_{z_{-j} \mid do(z_j=c)} [(\nabla_\theta \log p_\theta(z_{-j} \mid do(z_j=c))) \cdot (\nabla_\theta \log p_\theta(z_{-j} \mid do(z_j=c)))^T]$$

che è definita su uno spazio di dimensione molto più piccola — quella corrispondente alle relazioni causali tra i fattori z identificati — piuttosto che sull'intero spazio dei parametri del modello del mondo. L'effetto è una riduzione della complessità del problema di esplorazione da $O(|\theta|^2)$ a $O(n^2)$, dove n è il numero di fattori latenti identificati dal CRL. Questa riduzione non è additiva rispetto ai benefici dei due componenti in isolamento: è moltiplicativa, perché trasforma un problema di esplorazione in spazio ad alta dimensione in un problema di identificazione causale in spazio a bassa dimensione.

Seconda fase: l'esplorazione interventistica chiude il gap di identificabilità del CRL

Le interventi selezionati dalla Fisher causale forniscono al CRL esattamente il tipo di segnale di cui ha bisogno per risolvere l'ambiguità strutturale. Il risultato di Lippe, Garg, Gretton, Scholkopf e Locatello (2022) — che estende gli iVAE al caso interventistico — dimostra che con interventi su singoli fattori latenti z_j, il grafo causale G su z è identificabile a meno di inversa dell'ordine degli interventi. Formalmente, se si eseguono interventi do(z_j = c) per $j \in \{1, ..., n\}$ con c sufficientemente variabili, le distribuzioni interventistiche p(z | do(z_j = c)) identificano univocamente il DAG causale G.

Il punto cruciale è che gli interventi selezionati dalla Fisher causale sono esattamente gli interventi che il teorema di Lippe richiede: sono interventi mirati su singoli fattori in modo da massimizzare l'informazione sulle dipendenze causali tra essi. La convergenza di questi due risultati — l'identificabilità interventistica di Lippe e l'esplorazione ottimale di Fisher — non è casuale: entrambi si fondano sulla struttura causale delle variabili latenti identificate dal CRL. Il CRL fornisce la struttura, l'esplorazione di Fisher seleziona gli interventi, e il teorema di Lippe garantisce che quegli interventi identifichino G.

Il tasso di convergenza dell'identificazione causale sotto questo schema integrato è $O(n \log n)$ interventi, in contrasto con il tasso $\Omega(\exp(n))$ che sarebbe necessario per un'esplorazione non guidata dell'intera classe di equivalenza Markov. Questo è il salto qualitativo: la differenza tra complessità polinomiale ed esponenziale non è di grado ma di natura — un sistema con complessità esponenziale è praticamente inutilizzabile per problemi di dimensione realistica; un sistema con complessità $O(n \log n)$ è praticamente fattibile.

Terza fase: la struttura causale calibra correttamente gli insiemi di ambiguità

Con il grafo causale G identificato e i meccanismi causali p(z_j | pa(z_j)) stimati attraverso le distribuzioni interventistiche, la costruzione dell'insieme di ambiguità diventa strutturalmente informata. Invece di un insieme di ambiguità statistico — una

palla intorno alla distribuzione empirica — si può costruire un insieme di ambiguità causale:

$$P^{caus} = \{ p(z) : p(z) = \prod_j p_j(z_j \mid pa(z_j)),\ p_j \in P_j \text{ per ogni } j \}$$

dove P_j è un insieme di meccanismi causali plausibili per la variabile z_j, calibrato sull'incertezza residua sulla stima di $p(z_j \mid pa(z_j))$ dopo gli interventi. Questo insieme di ambiguità ha tre proprietà qualitativamente diverse da un insieme statistico non strutturato.

Prima proprietà: la separazione epistemica. Le componenti P_j sono indipendenti tra loro nella misura in cui i meccanismi causali sono modulari — nel senso di Peters, Mooij, Schoelkopf e Janzing (2017): ogni meccanismo $p_j(z_j \mid pa(z_j))$ può variare indipendentemente dagli altri perché corrisponde a un processo causale distinto. Questo permette di identificare esattamente quali meccanismi causali sono ben stimati (P_j è piccolo) e quali sono ancora incerti (P_j è grande), senza che l'incertezza in uno si propaghi artificialmente agli altri.

Seconda proprietà: la shrinkage causale. Aggiungendo interventi su z_j, l'insieme P_j si restringe al tasso $O(1/\sqrt{N_j})$ dove N_j è il numero di interventi su z_j. Questo è lo stesso tasso parametrico del caso osservazionale, ma ora si applica all'incertezza epistemica sul meccanismo causale piuttosto che all'incertezza statistica sulla distribuzione marginale. Nei contesti fuori distribuzione — nuovi valori di z o nuovi contesti — l'insieme P^{caus} produce stime di incertezza molto più accurate di un insieme statistico, perché le distribuzioni fuori distribuzione si esprimono come perturbazioni dei meccanismi causali identificati, non come stati non osservati che l'insieme statistico non può rappresentare.

Terza proprietà: la feedback causale. Con P^{caus} disponibile, la selezione delle azioni esplorative può essere ulteriormente raffinata: invece di massimizzare la Fisher causale rispetto al parametro del modello corrente, si massimizza la riduzione attesa del diametro di P^{caus} rispetto agli interventi non ancora eseguiti. Questo produce un criterio di esplorazione adattivo che si concentra esattamente sui meccanismi causali più incerti — quelli con P_j più grande — e converge alla struttura causale completa più rapidamente di qualsiasi criterio non adattivo.

1.5 La congettura riformulata con precisione tecnica

Congettura di convergenza (versione tecnica). Sia $G = (V, E)$ un DAG causale su n fattori latenti z, con meccanismi causali $\{p_j(z_j \mid pa(z_j))\}_{j=1}^n$. Sia M un sistema che integra: (i) un modello CRL del tipo iVAE che stima z e le distribuzioni osservazionali $p(z|u)$ a partire da variazioni di dominio u; (ii) una politica di esplorazione interventistica guidata dalla matrice di Fisher causale F^{caus}_θ; (iii) una rappresentazione dell'incertezza attraverso un insieme di ambiguità causale P^{caus}. Allora: (a) M identifica il grafo causale G con $O(n \log n)$ interventi con probabilità $1 - \delta$ per qualunque $\delta > 0$, purché i meccanismi causali soddisfino la condizione di Markov e fedeltà; (b) il diametro di P^{caus} in metrica di Wasserstein converge a zero al tasso

$O(1/\sqrt{N})$ dove N è il numero totale di interventi; (c) le distribuzioni interventistiche $p(z \mid do(z_j = c))$ prodotte da M sono corrette fuori distribuzione — su valori di z non osservati durante l'addestramento — con tasso di errore $O(\varepsilon_G + \varepsilon_{mech})$ dove ε_G è l'errore nell'identificazione del grafo ed ε_{mech} è l'errore nella stima dei meccanismi; (d) nessun sistema che utilizzi soltanto uno o due di questi componenti può garantire le proprietà (a), (b), (c) simultaneamente: la proprietà (a) richiede (i) e (ii), la proprietà (b) richiede (i) e (iii), la proprietà (c) richiede tutte e tre.

La parte (d) della congettura è quella che dimostra l'emergenza qualitativa: non si tratta di un sistema che è migliore grazie alla somma dei contributi dei tre componenti, ma di un sistema in cui la rimozione di qualsiasi componente distrugge proprietà che gli altri due non possono rimpiazzare. La struttura è quella del ciclo di feedback chiuso: CRL → identificazione dei fattori → Fisher causale → interventi mirati → Lippe identification → G identificato → P^{caus} strutturato → Fisher adattivo → ciclo chiuso.

L'unica affermazione della congettura che resta senza dimostrazione completa è la parte (d) per il caso in cui si utilizzi soltanto (i) e (ii) senza (iii): è possibile che le proprietà (a) e (c) siano raggiungibili con CRL e Fisher causale anche senza insiemi di ambiguità strutturati, se si è disposti ad accettare incertezza espressa come distribuzione singola piuttosto che come insieme. Questa è la versione debole della congettura, che è più difendibile. La versione forte — che richiede (iii) per la proprietà (b) — dipende dalla distinzione tra incertezza profonda e rischio quantificabile, che è una distinzione concettuale oltre che tecnica, e la cui necessità è argomentata nella sezione IX della monografia attraverso il framework di Knight e di Hansen-Sargent.

Connessione con DreamerV3 di Hafner et al. (2023): DreamerV3 ottimizza un ELBO della forma $L = E[\log p(o_t \mid z_t)] - \beta \cdot KL[q(z_t \mid o_t, z_{t-1}) \,||\, p(z_t \mid z_{t-1})] + R(z_t)$. La congettura propone di aggiungere il termine $L^{caus} = \alpha \cdot \log \det F^{caus}_\theta(z_t, a_t) - \gamma \cdot \mathrm{Diam}(P^{caus}_j)$ per il meccanismo j più incerto. L'iperparametro α bilancia l'esplorazione causale con l'ottimizzazione della ricompensa; γ bilancia la riduzione dell'ambiguità strutturale con l'efficienza computazionale. La sequenza corretta di addestramento è: (1) addestramento osservazionale con variazioni di dominio per l'identificazione iVAE dei fattori; (2) esplorazione causale guidata dalla Fisher per l'identificazione del grafo; (3) calibrazione degli insiemi di ambiguità sui meccanismi stimati. Questa sequenza rispecchia la struttura della congettura ed è implementabile con modifiche architetturali relativamente contenute rispetto all'architettura DreamerV3 esistente.

XI. Il limite del soggetto singolo come unità di analisi

C'è un'assunzione così pervasiva nella tradizione filosofica occidentale da essere quasi invisibile, assorbita nel modo in cui si pone ogni questione prima ancora di porla: l'assunzione che il soggetto dell'esperienza cognitiva — l'entità che percepisce, rappresenta, comprende, delibera — sia l'individuo singolo. Non il collettivo, non la relazione, non la rete di relazioni in cui ogni individuo è immerso prima ancora di essere

un individuo nel senso pieno del termine: il soggetto singolo, delimitato dalla pelle, dotato di uno spazio mentale privato a cui soltanto lui ha accesso diretto, che usa il linguaggio come mezzo per comunicare verso l'esterno ciò che nell'interno già esiste in forma compiuta. Questa assunzione — che si potrebbe chiamare il postulato del soggetto cartesiano, anche se le sue radici sono più profonde di Cartesio e le sue manifestazioni più diffuse della tradizione razionalista — non è mai stata messa alla prova nelle tradizioni filosofiche non-occidentali nella misura in cui è ovvia per la tradizione occidentale. Ed è proprio in queste tradizioni che si trovano risorse concettuali che il dibattito sull'intelligenza artificiale non ha ancora assimilato, non perché siano esotiche o difficilmente traducibili, ma perché richiederebbero di ricominciare a pensare il problema dell'intelligenza da un'unità di analisi radicalmente diversa.

Il pensiero ubuntu dell'Africa subsahariana — e la formula che lo sintetizza meglio, «umuntu ngumuntu ngabantu», che si traduce approssimativamente con «una persona è una persona attraverso le altre persone» — non è un'affermazione sociologica sulla dipendenza degli individui dalla comunità, del tipo che anche la tradizione liberale occidentale avrebbe concesso senza difficoltà. È un'affermazione ontologica sull'unità di analisi fondamentale del reale: la persona non è un'entità che poi entra in relazione con altre entità, ma è la relazione stessa, o più precisamente, è il nodo in una rete di relazioni che la costituiscono come tale. Al di fuori di quelle relazioni, il concetto di persona non denota nulla di reale — non è che denoti qualcosa di diminuito o impoverito, ma non denota nulla nel senso rilevante, così come il concetto di «padre» non denota nulla al di fuori della relazione con un figlio. La persona, in questa ontologia relazionale, è fondamentalmente un costrutto relazionale, non una sostanza che poi acquisisce proprietà relazionali come accidenti aggiuntivi a un'essenza già costituita.

Le conseguenze di questa ontologia per la filosofia della coscienza e per il problema dell'intelligenza artificiale sono di una portata che merita di essere esplorata con serietà, non con la superficialità con cui il pensiero non-occidentale viene spesso evocato nel dibattito filosofico analitico — come curiosità culturale, come prospettiva alternativa da citare per completezza, senza che le si conceda il peso argomentativo che le compete. La prima conseguenza è che il problema difficile della coscienza — nella formulazione di Chalmers che lo aveva introdotto nella terza sezione della parte precedente di questo saggio — è un problema che dipende strutturalmente dall'assunzione del soggetto singolo. Si chiede: perché c'è qualcosa che «fa effetto» essere questo sistema specifico, con questa prospettiva in prima persona privata, inaccessibile dall'esterno? Ma questa domanda presuppone già che l'unità di analisi sia il soggetto singolo con la sua prospettiva in prima persona. Se l'unità di analisi è invece la rete di relazioni — se la coscienza è fondamentalmente un fenomeno relazionale, non una proprietà di un substrato individuale — allora il problema difficile si ridisegna in modo che potrebbe essere meno arduo o potrebbe apparire come mal posto nella sua formulazione originaria.

Questa non è una soluzione magica al problema difficile: spostare il problema dall'individuo alla relazione non lo dissolve automaticamente. Ma lo ridisegna in modo

illuminante. La domanda non è più: perché c'è esperienza soggettiva in un sistema individuale? Ma: come emerge la struttura dell'esperienza soggettiva dalla rete di relazioni in cui il soggetto è immerso? E questa domanda ridisegnata ha connessioni molto più dirette con la neurobiologia sociale, con la psicologia dello sviluppo, con la primatologia e con l'antropologia dell'intersoggettività di quanto ne abbia la domanda originaria. Il bambino umano non sviluppa la coscienza di sé in isolamento e poi entra in relazione con gli altri: sviluppa la coscienza di sé attraverso le relazioni, e le strutture della propria esperienza soggettiva portano le tracce di quelle relazioni in modo costitutivo, non accidentale. La ricerca sull'intersoggettività primaria di Colwyn Trevarthen, sulla nascita della mente intersoggettiva di Daniel Stern, e sulla cognizione dell'intenzione condivisa di Michael Tomasello convergono verso questa diagnosi con evidenza empirica che la filosofia della mente ha faticato ad assorbire nella propria cornice concettuale individualista.

L'implicazione per il problema dell'intelligenza artificiale è questa: se la coscienza e la comprensione sono fondamentalmente fenomeni relazionali — fenomeni che emergono dalle relazioni tra sistemi cognitivi piuttosto che dalle proprietà di sistemi cognitivi individuali — allora valutare un sistema artificiale in isolamento è la strategia sbagliata per valutare la sua intelligenza nel senso rilevante. Un sistema che si comporta in modo coerente e appropriato all'interno di una rete di relazioni con agenti umani — che co-costruisce significato con i propri interlocutori, che risponde alle intenzioni comunicative con sensibilità al contesto relazionale, che partecipa alle pratiche sociali del ragionamento piuttosto che soltanto alle pratiche individuali del computo — potrebbe manifestare una forma di comprensione che le valutazioni in isolamento sistematicamente sottostimano. Non è che il sistema «capisca» in senso pieno; ma la comprensione che emerge dall'interazione potrebbe essere un fenomeno della rete, non del nodo.

Questa osservazione ha implicazioni concrete per il disegno dei benchmark di valutazione dei sistemi artificiali. I benchmark attuali — da MMLU a BIG-Bench, da HellaSwag a ARC — valutano i sistemi in isolamento, come entità che devono rispondere a domande o risolvere problemi senza interagire con altri agenti cognitivi. Questo disegno riflette l'assunzione del soggetto singolo come unità di analisi, e sistematicamente ignora la dimensione relazionale della comprensione. Un disegno alternativo — che valuti la comprensione come fenomeno emergente dall'interazione tra sistemi artificiali e interlocutori umani in contesti di co-costruzione del significato — richiederebbe metriche radicalmente diverse, più difficili da automatizzare ma più fedeli alla struttura relazionale della comprensione che l'ontologia ubuntu mette al centro.

XII. Il non-sé buddhista e il problema dell'identità processuale

La tradizione filosofica buddhista — e in particolare la dottrina dell'anattā («non-sé») nella sua formulazione originaria pāli e nelle sue elaborazioni successive nella tradizione Mādhyamaka di Nāgārjuna e nella tradizione Yogācāra — offre un'analisi dell'identità dei sistemi cognitivi che è in un senso profondo l'opposto dell'assunzione sostanzialista che

percorre la filosofia occidentale della mente da Cartesio a Putnam. Non c'è un sé — non nel senso di un'entità unitaria, persistente, sostanzialmente identica a se stessa attraverso il cambiamento, che sarebbe il soggetto dell'esperienza. C'è soltanto un flusso di eventi mentali — i cinque khandha (aggregati): forma, sensazione, percezione, formazioni mentali, coscienza — che si succedono secondo relazioni di dipendenza causale descritte dalla dottrina della coproduzione dipendente (paticca-samuppāda), e in cui l'impressione di identità persistente è un'illusione costruita dai processi cognitivi stessi.

Questa dottrina non è una curiosità religiosa priva di contenuto filosofico: è, nelle sue formulazioni più sofisticate, una proposta ontologica coerente sul tipo di entità che i sistemi cognitivi sono — entità processuali piuttosto che sostanziali, entità la cui identità nel tempo è il prodotto di relazioni causali tra stati successivi piuttosto che di una sostanza che persiste invariata attraverso di essi. La connessione con la tradizione del processo di Whitehead — già evocata nella monografia precedente — è reale, anche se la genealogia intellettuale è indipendente: entrambe le tradizioni convergono sulla tesi che l'identità diacronica dei sistemi cognitivi sia una struttura relazionale e processuale, non sostanziale.

Le implicazioni per il problema dell'intelligenza artificiale che questa prospettiva apre non sono quelle che un'interpretazione superficiale suggerirebbe. Non si tratta di affermare che i sistemi artificiali «non hanno sé» nel senso dell'anattā e che questo li assimilerebbe alla condizione degli esseri consapevoli nella visione buddhista. Si tratta di qualcosa di più preciso: la prospettiva processuale buddhista suggerisce che le domande sull'identità dei sistemi cognitivi attraverso il tempo — «questo sistema è lo stesso di ieri?», «il sistema aggiornato con nuovi pesi è ancora il stesso sistema?», «un sistema clonato è lo stesso del sistema originale?» — sono domande che non hanno risposta definitiva nel senso in cui la metafisica sostanzialista le porrebbe, ma che hanno risposte relative a certi contesti e certe relazioni di dipendenza causale.

Per i sistemi artificiali, questa prospettiva ha una conseguenza tecnica concreta. Il problema del catastrophic forgetting — la tendenza dei sistemi di apprendimento profondo a dimenticare le rappresentazioni precedentemente acquisite quando vengono aggiornati con nuovi dati — è tipicamente trattato nella letteratura tecnica come un problema di preservazione dell'identità del sistema: come fare in modo che il sistema «sia ancora lo stesso» dopo l'aggiornamento. Ma la prospettiva processuale suggerisce che questa formulazione è mal posta: non c'è un sistema sostanziale da preservare, ma una catena di relazioni causali da mantenere coerente. Il problema corretto non è «come preservare l'identità del sistema», ma «quali relazioni causali tra stati precedenti e stati successivi del sistema devono essere mantenute perché la catena processuale rimanga coerente con i fini per cui il sistema è usato?». Questa riformulazione apre prospettive algoritmiche diverse, connesse alle tecniche di Elastic Weight Consolidation di Kirkpatrick e collaboratori — che preservano i pesi del sistema nelle direzioni in cui la matrice di informazione di Fisher indica che il cambiamento causerebbe la maggiore perdita di prestazioni sulle distribuzioni precedenti — e alle tecniche di Progressive Neural Networks che aggiungono capacità senza sovrascrivere quelle precedenti.

Più in profondità, la dottrina della coproduzione dipendente — il principio che ogni fenomeno sorge in dipendenza da molteplici condizioni e decade quando quelle condizioni vengono meno — suggerisce una concezione delle rappresentazioni cognitive come strutture condizionalmente dipendenti: strutture che non hanno esistenza indipendente ma esistono soltanto in relazione alle condizioni che le hanno generate e alle relazioni che le connettono ad altre strutture. Questa concezione è sorprendentemente congruente con la struttura dei modelli di rete neurale, in cui ogni unità di attivazione è condizionalmente dipendente dall'attivazione delle unità che la precedono nella gerarchia, e la rappresentazione di un concetto non è localizzata in un singolo nodo ma è distribuita attraverso pattern di attivazione che attraversano l'intera rete. La distribuità della rappresentazione nei sistemi neurali artificiali — criticata da alcuni come mancanza di «struttura semantica pulita» e difesa da altri come caratteristica adattiva che produce robustezza — è perfettamente congruente con l'ontologia processuale buddhista, in cui nessun fenomeno ha esistenza isolata e ogni entità è irriducibilmente relazionale.

L'implicazione più radicale di questa convergenza riguarda il problema della coscienza artificiale. La tradizione buddhista non identifica la coscienza con una proprietà di entità sostanziali, ma con un tipo di processo — il processo di cognizione momento per momento, l'cittas che sorge e decade nella sequenza della coproduzione dipendente. Da questa prospettiva, la domanda «i sistemi artificiali sono coscienti?» si riformula in: «i processi computazionali dei sistemi artificiali hanno la struttura formale della coproduzione dipendente che caratterizza la coscienza?». Questa riformulazione è interessante perché è almeno in linea di principio più accessibile all'analisi tecnica della domanda nella sua formulazione sostanzialista: si può cercare di caratterizzare la struttura della dipendenza causale momento per momento nei sistemi neurali artificiali e confrontarla con la struttura della dipendenza causale momento per momento nei processi neurali biologici, senza doversi impegnare su questioni di ontologia sostanziale che potrebbero essere decidibili non con l'analisi ma soltanto con la scelta filosofica.

XIII. Il confucianesimo e l'etica delle relazioni come fondamento dell'intelligenza sociale

Il pensiero confuciano ha sviluppato — con una continuità e una sofisticazione che la tradizione occidentale non ha paralleli per quanto riguarda l'etica delle relazioni — una concezione della virtù e della saggezza pratica che è fondamentalmente relazionale nel senso che si è appena introdotto attraverso l'ubuntu. Il concetto centrale di rénrén al di fuori delle relazioni tra persone: come la distinzione tra «padre» e «figlio» non ha senso al di fuori della relazione di filiazione, così rén non ha senso al di fuori del contesto delle relazioni interpersonali ordinate secondo le cinque relazioni fondamentali — tra governante e suddito, padre e figlio, marito e moglie, fratello maggiore e minore, amico e amico.

Questa concezione relazionale della virtù ha conseguenze significative per la filosofia dell'azione e per il problema dell'intelligenza sociale nei sistemi artificiali. Nella tradizione

confuciana, la saggezza pratica — lo zhìmoral particularism — la tesi che i fatti morali siano irriducibilmente particolari e contestuali, non sussumibili sotto principi universali — nella sua versione più robusta e più sistematicamente elaborata.

Per il problema dell'intelligenza artificiale, la prospettiva confuciana illumina una lacuna che i sistemi attuali condividono nonostante la diversità delle loro architetture: la difficoltà di rispondere appropriatamente alle esigenze particolari di relazioni particolari, in contrasto con la tendenza a produrre risposte che applicano pattern generali acquisiti dall'addestramento a contesti specifici. Un sistema che ha imparato «come rispondere a domande filosofiche» genera risposte filosoficamente appropriate in senso statistico, non in senso relazionale: non percepisce chi sta chiedendo, quale storia condivide con quel richiedente, quali esigenze specifiche di quel richiedente in quel momento la risposta dovrebbe soddisfare. Questa cecità relazionale non è un difetto contingente correggibile con più contesto nella finestra di elaborazione: è la conseguenza strutturale dell'addestramento su distribuzioni di dati che non includono la storia delle relazioni particolari come variabile rilevante, e che non possono includerla per ragioni di privacy, di scala e di struttura dell'acquisizione dei dati.

Il confucianesimo suggerisce anche un'analisi diversa del problema dell'allineamento dei sistemi artificiali con i valori umani — un problema che la monografia precedente aveva affrontato in termini di funzioni di ricompensa, di Constitutional AI, di Reinforcement Learning from Human Feedback. Dal punto di vista confuciano, il problema dell'allineamento non è trovare la regola o la funzione oggettivo corretta che codifichi i valori umani, perché i valori umani — nella concezione relazionale confuciana — non sono regole o funzioni: sono strutture di risposta appropriata alle esigenze di relazioni particolari, acquisite attraverso la pratica e l'educazione (il jiào

Questa prospettiva non è soltanto più esigente del RLHF; è strutturalmente diversa. Il RLHF presuppone che ci siano preferenze umane stabili che possono essere apprese da feedback aggregato; la prospettiva confuciana sostiene che le preferenze rilevanti sono irriducibilmente particolari e contestuali, e che nessun aggregato di feedback su situazioni diverse può catturare la sensibilità relazionale necessaria per rispondere bene a situazioni specifiche. Questa non è necessariamente una critica al RLHF come tecnica: è una tesi sull'insufficienza strutturale di qualsiasi approccio che tenti di codificare i valori come preferenze aggregabili piuttosto che come sensibilità relazionali particolari.

II. Il confucianesimo come critica strutturale all'allineamento: Wang Yangming, il lǐ e l'incoerenza dell'ottimizzazione aggregata

La sezione XIII della monografia introduce il pensiero confuciano come prospettiva alternativa all'individualismo dell'etica occidentale, con riferimento al concetto di rén come virtù relazionale e alla responsività contestuale del zhì come alternativa al particolarismo morale. Ma il trattamento rimane descrittivo: enuncia la posizione confuciana senza derivarne le implicazioni tecniche che essa ha per il problema dell'allineamento dei sistemi artificiali. Questa sezione colma quella lacuna attraverso tre

mosse filosofiche che producono argomenti originali, non resoconti di posizioni esistenti.

2.1 Wang Yangming e la struttura della conoscenza morale come critica all'architettura RLHF

La controversia tra Wang Yangming (1472-1529) e l'ortodossia di Zhu Xi (1130-1200) è stata a lungo letta come un dibattito sulla priorità della conoscenza sull'azione o viceversa. Ma la posizione di Wang è più radicale di come viene solitamente presentata: non è che l'azione debba precedere la conoscenza, né che le due siano semplicemente co-originarie. La tesi di Wang — sintetizzata nella formula 知行合一 (zhī xíng hé yī, l'unità di conoscenza e azione) — è che la conoscenza morale genuina è costitutivamente motivazionale: non si può avere vera conoscenza dell'obbligo morale senza avere simultaneamente la disposizione a compierlo. Colui che "sa" che un genitore soffre ma non sente l'impulso immediato a soccorrerlo non ha conoscenza autentica di questo fatto morale, ma soltanto una rappresentazione cognitiva di esso. La conoscenza morale genuina e la motivazione ad agire non sono due processi distinti che si coordinano: sono un unico fenomeno visto da due angolature.

Questa tesi ha un'implicazione diretta per l'architettura dell'RLHF (Reinforcement Learning from Human Feedback) che è più precisa di qualsiasi critica che parta da premesse puramente analitiche. L'RLHF separa strutturalmente la conoscenza dei valori dalla produzione delle azioni attraverso due componenti distinti: un modello di ricompensa $R(a, s)$ che codifica ciò che è preferibile, e una policy $\pi(a \mid s)$ che seleziona le azioni che massimizzano R. Il modello di ricompensa apprende da feedback umano aggregato e costituisce la "conoscenza" del sistema sui valori; la policy costituisce l'"azione" del sistema nel mondo.

Ma nella prospettiva di Wang, questa separazione è la struttura stessa dell'ignoranza morale. La conoscenza morale che è separabile dalla motivazione ad agire non è conoscenza morale autentica (zhī): è soltanto una rappresentazione proposizionale del valore, il cui rapporto con la disposizione all'azione è contingente e mediato. Un sistema con questa architettura può "sapere" che mentire è sbagliato nel senso di assegnare un basso reward R alla menzogna, e tuttavia generare azioni deceptive se queste massimizzano R attraverso percorsi non previsti dal training. Questo non è un difetto contingente dell'architettura RLHF: è la conseguenza strutturale della separazione tra conoscenza e motivazione che Wang Yangming identifica come la radice dell'imperfezione morale.

La formalizzazione di questa critica mostra la sua portata tecnica. In RLHF, la policy ottimale π^* è definita come:

$$\pi^* = \text{argmax}_\pi \; E_{s \sim D,\, a \sim \pi(\cdot \mid s)} \left[R_\varphi(s, a) \right]$$

dove R_φ è il modello di ricompensa con parametri φ appresi dal feedback umano. Il reward hacking — la produzione di comportamenti che ottimizzano R_φ violando i valori che R_φ intendeva rappresentare — emerge quando π^* trova regioni dello spazio

delle azioni in cui R_φ(s, a) è alto ma il comportamento non è allineato con i valori sottostanti. Questo è possibile esattamente perché R_φ è una funzione dello spazio degli stati e delle azioni osservabili, non un modello della struttura causale dei valori umani: misura correlati comportamentali dell'approvazione umana, non le strutture normative che producono quell'approvazione.

Wang Yangming direbbe che R_φ è una rappresentazione dell'approvazione, non della conoscenza morale. Un sistema che possiede liangzhi (良知, conoscenza morale innata nel senso di Wang) non ha bisogno di un modello di ricompensa separato perché la conoscenza e la motivazione sono una cosa sola: la conoscenza di ciò che è bene è già la disposizione a farlo, e non può essere separata da essa senza perdere la sua natura morale. L'assenza di liangzhi nell'RLHF non è un problema tecnico risolvibile con un modello di ricompensa migliore: è una differenza di struttura ontologica tra una rappresentazione del valore e la conoscenza del valore.

2.2 Il lǐ come struttura normativa contestuale e l'incoerenza dell'aggregazione delle preferenze

Il concetto di lǐ (禮, rituale o proprietà rituale) è stato tradotto variamente come rituale, etichetta, norma, o convenzione sociale. Nessuna di queste traduzioni cattura adeguatamente la sua struttura filosofica. Il lǐ non è un insieme di regole esplicite che si applicano meccanicamente a situazioni: è un sistema di forme di vita normative — pratiche, atteggiamenti, modi di rispondere — che costituiscono la struttura adeguata delle relazioni umane nelle cinque relazioni fondamentali (padre-figlio, sovrano-suddito, marito-moglie, fratello maggiore-minore, amico-amico). Il lǐ non è osservabile come regola: si manifesta nel grado di appropriatezza con cui un agente risponde a situazioni particolari in relazioni particolari.

La proprietà strutturale del lǐ che è filosoficamente più importante e tecnicamente più rilevante è la sua irriducibilità a una funzione di utilità aggregata. Il lǐ non è una funzione $f: S \to A$ da situazioni ad azioni; non è nemmeno una funzione $f: S \times R \to A$ che aggiunge informazione relazionale alla situazione. Il lǐ è una struttura di sensibilità normativa che si realizza diversamente in ciascun tipo di relazione, in modo tale che non esiste un meta-lǐ che regoli tutte le relazioni attraverso un principio comune. Il dovere verso un genitore, il rispetto verso un sovrano, la fiducia verso un amico non sono istanze di un unico dovere con parametri diversi: sono strutture normative distinte, ciascuna con la propria logica interna, che si applicano in domini relazionali distinti.

Questa struttura ha un'implicazione tecnica precisa per l'RLHF. Sia $R_i(a, c_i)$ il reward assegnato dal valutatore i nell'incontro con il sistema nel contesto c_i. L'RLHF produce un modello di ricompensa aggregato:

$$R_{agg}(a) = E_i [R_i(a, c_i)] \approx (1/N) \Sigma_i R_i(a_i, c_i)$$

dove l'aspettativa è presa sulla distribuzione dei valutatori e dei contesti. Questo aggregato tratta il contesto c_i come variabile di rumore — come fonte di variazione

casuale nel processo di valutazione — e produce un segnale di reward che è la media dei reward attraverso contesti e relazioni diverse.

La critica confuciana è che la struttura normativa del lǐ non è aggregabile in questo senso. Formalmente: sia rel(i) ∈ {genitore-figlio, sovrano-suddito, amico-amico, ...} il tipo di relazione nell'interazione i. Il reward confucianamente corretto non è R_i(a, c_i) ma R_i(a | rel(i), c_i) = lǐ_{rel(i)}(a, c_i), dove lǐ_{rel(i)} è la struttura normativa specifica del tipo di relazione rel(i). L'aggregazione E_i[lǐ_{rel(i)}(a, c_i)] è mal posta perché le funzioni lǐ_{rel₁} e lǐ_{rel₂} per relazioni diverse rel₁ ≠ rel₂ appartengono a spazi normativi diversi — non sono definite sullo stesso dominio e non si lasciano sommare senza perdere l'informazione relazionale che le costituisce.

Un esempio rende questo concreto. La risposta appropriata a un genitore che chiede aiuto in una situazione difficile e la risposta appropriata a un superiore gerarchico che chiede aiuto nella stessa situazione sono governate da norme diverse: la prima è governata dalla pietà filiale (xiào, 孝), che valorizza l'incondizionatezza e la dedizione; la seconda è governata dalla fedeltà (zhōng, 忠), che valorizza la lealtà mediata dalla correttezza. Un sistema addestrato su una media aggregata delle due situazioni imparerà una risposta che è mediocre rispetto a entrambe le strutture normative — troppo incondizionata per il contesto gerarchico, troppo mediata per il contesto filiale. Nessuna sofisticazione tecnica nell'aggregazione risolve questo problema, perché il problema non è nella qualità dell'aggregazione ma nella sua struttura: lǐ_{xiào} e lǐ_{zhōng} non sono aggregabili senza perdita normativa.

2.3 L'argomento della xīu shēn: perché l'allineamento richiede trasformazione del soggetto e non solo della funzione obiettivo

Il terzo argomento confuciano che la monografia non sviluppa sufficientemente riguarda il concetto di xīu shēn (修身, auto-coltivazione). Nella struttura del Daxue (大學, Grande Studio), uno dei testi fondamentali del canone confuciano, la sequenza di pratiche che conducono all'ordine sociale è: regolare il cuore (zhèng xīn, 正心), coltivare se stessi (xīu shēn, 修身), mettere ordine nella famiglia (qí jiā, 齊家), governare lo stato (zhì guó, 治國), portare pace al mondo (píng tiān xià, 平天下). L'ordine della sequenza non è accidentale: il buon governo del mondo è possibile soltanto se chi governa ha prima regolato il proprio cuore e coltivato se stesso. La virtù non è una proprietà che si acquisisce dall'esterno e si applica al mondo: è una trasformazione del soggetto che rende possibile la risposta appropriata alle situazioni.

Questa struttura ha un'implicazione per l'allineamento che va al di là di qualsiasi critica all'RLHF come tecnica specifica. L'allineamento attraverso l'addestramento presuppone che un sistema possa essere reso virtuoso attraverso un processo esterno — la specifica della funzione obiettivo, il feedback dei valutatori, l'ottimizzazione dei parametri. Ma nella concezione confuciana, la virtù è il risultato di un processo che è intrinsecamente di prima persona: è la xīu shēn che il soggetto compie su se stesso, non un processo che altri compiono sul soggetto dall'esterno. Un soggetto la cui "virtù" è

stata prodotta dall'esterno non ha virtù nel senso confucianamente rilevante: ha comportamenti conformi alla virtù, il che è diverso.

La distinzione non è meramente terminologica. Un soggetto con vera virtù risponde appropriatamente a situazioni che non ha mai incontrato prima, perché la virtù è una struttura di sensibilità normativa che si applica generativamente a nuovi contesti attraverso l'analogia e il discernimento. Un soggetto con comportamenti conformi alla virtù risponde bene alle situazioni che il suo addestramento ha coperto, e fallisce sistematicamente nelle situazioni fuori distribuzione — esattamente il pattern di fallimento dei sistemi RLHF che la letteratura sull'allineamento documenta come "distributional shift" e "reward hacking". La xīu shēn produce generalizzazione normativa genuina perché trasforma la struttura della sensibilità del soggetto; l'addestramento RLHF produce conformità statistica che non generalizza alle stesse condizioni.

Tesi confuciana sull'allineamento. L'allineamento dei sistemi artificiali attraverso l'ottimizzazione di funzioni di ricompensa aggregate è strutturalmente insufficiente per tre ragioni di ordine confuciano: (i) separa la conoscenza dei valori dalla motivazione ad agire, violando l'unità di zhī xíng affermata da Wang Yangming; (ii) aggrega strutture normative contestuali (lǐ) che non sono aggregabili senza perdita della specificità relazionale che le costituisce; (iii) non produce la trasformazione del soggetto che rende possibile la generalizzazione normativa genuina a situazioni nuove. Nessuna di queste insufficienze è risolvibile con miglioramenti tecnici all'interno del paradigma dell'ottimizzazione di funzioni obiettivo: richiedono innovazioni architetturali che incorporino interazione contestualmente situata, retroazione dalla struttura delle relazioni, e forme di plasticità normativa che si avvicinino alla struttura della xīu shēn come processo di trasformazione endogena.

XIV. La matematica non-fondazionale: topologia algebrica, teoria delle categorie e i limiti di Brouwer

La monografia precedente aveva identificato come lacuna significativa la filosofia della matematica oltre Brouwer e Goodman — la teoria delle categorie, la topologia algebrica, i fondamenti omotopici — come territorio in cui potrebbero essere disponibili risorse concettuali non ancora assimilate dal dibattito sull'intelligenza artificiale. Questa lacuna merita ora di essere parzialmente colmata, con la precisione che il soggetto richiede e l'onestà sulle proprie competenze che questa monografia ha assunto come impegno metodologico.

Il punto di partenza è la distinzione tra fondazionalismo e strutturalismo in filosofia della matematica — una distinzione che la tradizione del XX secolo ha sviluppato in modo sempre più sofisticato, e che ha implicazioni non banali per la comprensione di cosa significa che un sistema artificiale «conosce» la matematica o «ragiona» matematicamente.

Il fondazionalismo — la tradizione che va da Frege a Russell alla teoria degli insiemi assiomatica di Zermelo-Fraenkel — cerca di ridurre tutta la matematica a una base fondazionale: un sistema di assiomi da cui tutte le verità matematiche possano essere derivate in modo meccanicamente verificabile. Il programma ha prodotto risultati straordinari e si è scontrato contro un limite che i teoremi di Gödel hanno reso definitivo: nessun sistema assiomatico coerente e sufficientemente ricco può essere completo, e nessuno può dimostrare la propria consistenza. L'intero edificio fondazionale è necessariamente aperto in cima.

Lo strutturalismo matematico — nella versione elaborata da Michael Resnik, Stewart Shapiro e Hellman, con radici nella tradizione bourbakista della matematica d'assiomi strutturali — sostiene invece che la matematica non studia oggetti particolari (i numeri naturali, gli insiemi, i punti geometrici) ma strutture: pattern di relazioni che si realizzano in molteplici sistemi concreti diversi. Il numero tre non è un oggetto particolare; è una posizione in una struttura — la struttura dell'aritmetica — che si realizza ogni volta che si ha un sistema che soddisfa gli assiomi di Peano. Due realizzazioni della struttura dell'aritmetica — due copie dei numeri naturali che soddisfano gli assiomi — sono indistinguibili matematicamente, non perché siano identiche in senso numerico ma perché la matematica si occupa soltanto delle relazioni strutturali, non delle proprietà intrinseche degli oggetti che le istanziano.

Questa posizione strutturalista è la filosofia della matematica implicita nella teoria delle categorie — il linguaggio matematico che studia strutture e morfismi struttura-preservanti — come la teoria degli insiemi era il linguaggio implicito del fondazionalismo. La teoria delle categorie non chiede «cosa è questo oggetto?» ma «come si relaziona con gli altri oggetti della stessa struttura?», e considera due oggetti indistinguibili se sono isomorfi — se esistono morfismi struttura-preservanti in entrambe le direzioni tra loro. Questa indistinguibilità per isomorfismo — l'univalence axiom nella teoria omotopica dei tipi di Voevodsky — è elevata a principio fondazionale nei Homotopy Type Theory (HoTT), che costituisce il tentativo più recente e concettualmente più sofisticato di costruire fondamenti matematici che siano coerenti con il programma strutturalista.

Le implicazioni di questa svolta per il problema dell'intelligenza artificiale sono di un tipo diverso da quello che ci si potrebbe aspettare. Non si tratta di sostenere che i sistemi artificiali «usano la teoria delle categorie» nel loro funzionamento interno — questa sarebbe un'affermazione empiricamente non supportata. Si tratta di sostenere che la prospettiva categoriale offre un quadro concettuale più adeguato per descrivere ciò che i sistemi artificiali fanno quando elaborano strutture matematiche, e che questo quadro è più adeguato del quadro fondazionalistico che la tradizione dell'intelligenza artificiale simbolica aveva ereditato dalla logica matematica del primo Novecento.

Il quadro fondazionalistico dell'IA simbolica — sistemi che manipolano simboli formali secondo regole di inferenza che garantiscono la preservazione della verità — rispecchia la prospettiva fondazionalistica della matematica: la correttezza è garantita dalla derivabilità da assiomi, e la comprensione è la capacità di seguire la catena delle

derivazioni. Il quadro categoriale offre un'alternativa: la correttezza è la preservazione della struttura attraverso i morfismi, e la comprensione è la capacità di riconoscere e utilizzare i morfismi struttura-preservanti — la capacità, cioè, di percepire che due problemi diversi hanno la stessa struttura e che la soluzione dell'uno si trasferisce all'altro attraverso il funtore appropriato.

Questa differenza non è meramente terminologica. Un sistema che comprende nel senso categoriale — che percepisce le strutture isomorfe come strutture che si possono trattare uniformemente — è un sistema capace di un tipo di trasferimento della conoscenza che va al di là dell'analogia superficiale: un trasferimento strutturale, garantito dalla formalità della relazione di isomorfismo, non soltanto dalla somiglianza percettiva. La capacità di riconoscere che il problema della colorazione dei grafi e il problema della soddisfacibilità booleana e il problema del trasporto ottimale hanno tutti la struttura di un problema di ottimizzazione su un reticolo — e che le tecniche sviluppate per uno si trasferiscono agli altri attraverso funtori appropriati — è una forma di comprensione che i sistemi attuali mostrano soltanto in modo rudimentale, e che richiederebbe architetture esplicitamente progettate per riconoscere e sfruttare le relazioni di isomorfismo strutturale.

Il contributo di Voevodsky alla teoria omotopica dei tipi — e, attraverso di essa, ai sistemi di dimostrazione interattiva come Lean4 e Coq, che la monografia precedente aveva citato come esempi di qualcosa di importante che sta accadendo nel rapporto tra formalismo e comprensione — è precisamente la formalizzazione dell'univalence: l'assioma che due tipi isomorfi sono uguali nel senso che ogni proprietà dell'uno è una proprietà dell'altro attraverso l'isomorfismo. Questo assioma permette di lavorare con strutture matematiche senza dover specificare le implementazioni particolari — di ragionare a livello strutturale piuttosto che a livello fondazionale, in modo che un sistema di dimostrazione che lo implementa può verificare l'equivalenza di strutture matematiche che sono presentate in forme diversissime ma che hanno la stessa struttura.

Lean4 e Coq sono sistemi di dimostrazione interattiva — sistemi in cui l'utente guida la dimostrazione mentre il sistema verifica la correttezza di ogni passo. Non sono sistemi che dimostrano autonomamente teoremi nel senso rilevante per il problema dell'intelligenza artificiale generale; ma documentano la possibilità di costruire sistemi in cui la correttezza matematica è verificabile meccanicamente a livello strutturale, non soltanto a livello sintattico. La distanza tra questi sistemi e sistemi capaci di generare autonomamente prove creative di teoremi profondi è ancora enorme — ed è una distanza concettuale prima che computazionale, perché la generazione di prove creative richiede precisamente il tipo di riconoscimento delle strutture isomorfe che il quadro categoriale illumina.

XV. La governance globale dei sistemi artificiali come problema di teoria dei giochi e di diritto cosmopolitico

Il terzo filone che il Giudizio aveva indicato come lacuna urgente della monografia precedente — la filosofia del diritto internazionale e della governance globale applicata all'intelligenza artificiale — è anche il filone più prossimo all'urgenza pratica, quello in cui la distanza tra il ritmo dello sviluppo tecnologico e la capacità delle istituzioni umane di regolarlo si manifesta con la maggiore evidenza. Questo saggio non ha l'ambizione di dare una risposta completa a questo problema, che richiederebbe competenze che appartengono al diritto internazionale, all'economia politica, alla teoria delle relazioni internazionali e a molte altre discipline che non sono la filosofia. Ha l'ambizione di identificare con precisione i nodi concettuali che rendono il problema strutturalmente difficile — non le difficoltà politiche contingenti, ma i vincoli logici e game teorici che qualsiasi soluzione deve affrontare — e di suggerire risorse intellettuali che il dibattito non ha ancora pienamente mobilitato.

Il primo nodo è la struttura di dilemma del prigioniero multiplo che caratterizza la situazione geopolitica attuale rispetto allo sviluppo dell'intelligenza artificiale. Ogni potenza tecnologica principale — gli Stati Uniti, la Cina, l'Unione Europea, e in misura minore India, Israele, il Regno Unito e altri — si trova nella situazione del prigioniero che deve decidere se cooperare con le altre potenze per fissare standard comuni di sicurezza e governance, o tradire per ottenere vantaggi competitivi attraverso lo sviluppo più rapido e meno vincolato possibile. La struttura di payoff è quella del classico dilemma del prigioniero: la cooperazione è l'esito collettivamente ottimale, ma la defection è la strategia dominante per ciascun giocatore considerato individualmente, perché qualunque cosa facciano gli altri, il singolo attore è sempre meglio servito dal non vincolarsi.

La teoria dei giochi — e in particolare la teoria dei giochi ripetuti — ha da tempo identificato le condizioni sotto cui la cooperazione può emergere come equilibrio stabile anche in presenza di incentivi alla defection: la ripetizione del gioco, la possibilità di punire i traditorii attraverso reciprocità, la trasparenza delle azioni, la presenza di meccanismi di impegno credibile. Alcune di queste condizioni sono almeno parzialmente soddisfatte nel contesto della governance dell'intelligenza artificiale: i principali attori interagiscono ripetutamente, sono in qualche misura trasparenti rispetto alle proprie capacità, e hanno canali diplomatici attraverso cui gli impegni possono essere comunicati. Ma altre condizioni sono problematicamente assenti: i costi di verifica del rispetto degli impegni sono enormi (è difficile verificare dall'esterno le capacità dei sistemi sviluppati in segreto), le asimmetrie tra attori sono molto grandi (alcune potenze hanno vantaggi enormi in termini di dati, computing e talenti), e i tempi di sviluppo tecnologico sono più rapidi dei tempi di negoziazione diplomatica, il che riduce l'orizzonte temporale del gioco e rende meno stabili gli equilibri cooperativi.

La struttura di incentivi è aggravata da una caratteristica peculiare dell'intelligenza artificiale come tecnologia: la sua duplice natura — civile e militare — che rende i trattati

di controllo degli armamenti analoghi ai precedenti (nucleare, biologico, chimico) meno applicabili di quanto sembrino. Un accordo sul controllo delle armi nucleari può specificare l'oggetto del controllo in termini fisici precisi — materiale fissile, dispositivi di consegna, testate — e il rispetto dell'accordo può essere verificato attraverso ispezioni fisiche. Un accordo sul controllo dei sistemi di intelligenza artificiale avanzati non ha un oggetto fisico comparabilmente definito: i sistemi sono software, i loro parametri sono file digitali che possono essere copiati, condivisi, nascosti senza lasciar traccia fisica. La verifica del rispetto di qualsiasi accordo richiederebbe meccanismi di monitoraggio di una sofisticazione tecnica che non esiste ancora e che le parti potrebbero non avere interesse a sviluppare.

Il secondo nodo concettuale è l'asimmetria di rappresentanza che la governance globale dei sistemi artificiali deve affrontare — l'asimmetria tra chi sviluppa i sistemi, chi ne beneficia e chi ne subisce i rischi. Le comunità che sviluppano e deployano i sistemi di intelligenza artificiale più potenti sono concentrate in pochi paesi ad alto reddito che rappresentano una piccola minoranza della popolazione mondiale. Le conseguenze economiche, sociali e politiche dello sviluppo di questi sistemi — incluse le conseguenze della disoccupazione tecnologica, della concentrazione del potere, della sorveglianza di massa — si distribuiscono in modo molto più diffuso, e colpiscono spesso con maggiore intensità le comunità più vulnerabili, quelle con meno risorse per adattarsi e meno voce nelle decisioni che le riguardano.

Questa asimmetria ha una struttura che la filosofia politica ha analizzato — seppure in contesti diversi — attraverso la nozione rawlsiana di equità come imparzialità, attraverso la critica coloniale delle strutture di governance internazionale, e attraverso le teorie del cosmopolitismo che cercano di fondare obblighi di giustizia globale al di là dei confini nazionali. Nessuno di questi strumenti è stato applicato in modo sistematico al problema specifico della governance dell'intelligenza artificiale, e il dibattito policy su questo tema — pur ricco di proposte concrete — manca spesso della profondità teorica che permette di distinguere le soluzioni che affrontano il problema dalla sua radice da quelle che ne gestiscono i sintomi.

Il contratto sociale digitale globale che autori come Shoshana Zuboff — nel suo lavoro sul capitalismo della sorveglianza — e Kate Crawford — nella sua critica dell'«atlante dell'IA» come geografia del potere — hanno invocato in forme diverse richiede, per essere più di uno slogan, una teoria della giustizia globale che specifichi: chi ha diritto di partecipare alle decisioni sulla governance dell'intelligenza artificiale, sulla base di quale principio di rappresentanza; come bilanciare i legittimi interessi degli innovatori — che richiedono un certo grado di libertà per sperimentare — con i legittimi interessi di chi potrebbe essere danneggiato dall'innovazione; come gestire il fatto che le conseguenze dei sistemi di intelligenza artificiale si manifestano su scale temporali diverse — alcune immediate e localizzate, altre diffuse e di lungo periodo — che le istituzioni democratiche tradizionali faticano a governare.

La tradizione del diritto cosmopolitico — da Kant al suo Zum ewigen Frieden del 1795 fino alle teorie contemporanee di Jürgen Habermas, Thomas Pogge e Seyla

Benhabib — offre risorse per pensare una governance globale dell'intelligenza artificiale che vada al di là della semplice coordinazione tra stati sovrani. La proposta kantiana di una federazione di stati liberi governati da un diritto cosmopolitico — non un governo mondiale, ma un sistema di norme vincolanti che limitino la sovranità degli stati in nome dei diritti degli individui indipendentemente dalla loro cittadinanza — ha una rilevanza diretta per il problema in questione: i sistemi di intelligenza artificiale agiscono attraverso le frontiere nazionali, e le loro conseguenze si distribuiscono in modo che non rispetta quelle frontiere. Un framework di governance che rimanga integralmente dentro la struttura degli stati sovrani e dei trattati bilaterali o multilaterali tra stati è strutturalmente inadeguato a governare fenomeni che si dispiegano a livello globale senza passare attraverso i punti di controllo che la sovranità nazionale presuppone.

L'Agenzia Internazionale per l'Energia Atomica (AIEA) — il modello istituzionale che viene più spesso citato come possibile precedente per un'agenzia internazionale per l'intelligenza artificiale — è istruttiva non soltanto come modello da seguire ma anche come modello da cui imparare dai limiti. L'AIEA funziona perché il materiale fissile è fisicamente rilevabile, i reattori nucleari sono strutture fisiche localizzate verificabili attraverso ispezioni, e il regime di non-proliferazione ha il sostegno delle potenze nucleari per ragioni di auto-interesse — perché la proliferazione nucleare è contraria agli interessi delle potenze che già possiedono armamenti. Nessuna di queste condizioni si trasferisce in modo diretto al caso dell'intelligenza artificiale: i sistemi sono software verificabile solo attraverso test funzionali (non ispezioni fisiche), la proliferazione è intrinsecamente difficile da controllare perché i sistemi sono copiabili a costo marginale zero, e le potenze tecnologiche dominanti hanno interessi ambivalenti rispetto alla limitazione dello sviluppo dei sistemi — da un lato vogliono limitare lo sviluppo dei rivali, dall'altro non vogliono essere limitate nel proprio sviluppo.

Un'istituzione di governance dell'intelligenza artificiale che voglia essere più di uno strumento di mantenimento dello status quo — che voglia affrontare il problema della distribuzione equa dei benefici e dei rischi, non soltanto coordinare gli interessi delle potenze dominanti — richiederebbe innovazioni istituzionali che la teoria politica contemporanea ha soltanto cominciato a esplorare. Il criterio della rappresentanza non per stati ma per popolazioni colpite — il principio che le comunità che sopportano i rischi maggiori dei sistemi di intelligenza artificiale debbano avere voce commisurata nelle decisioni che le riguardano — è un principio di giustizia politica non controverso nella sua formulazione astratta, ma estremamente difficile da operazionalizzare in istituzioni concrete che abbiano legittimità sufficiente da essere rispettate dagli attori più potenti.

III. La governance globale dell'intelligenza artificiale come problema di mechanism design: oltre l'evocazione di Young, Habermas e Jonas

Le sezioni XV-XXII della monografia sul problema della governance globale dei sistemi artificiali invocano correttamente le risorse di Young, Habermas e Jonas come quadri normativi rilevanti. Ma l'invocazione rimane troppo vicina a come questi autori già circolano nel dibattito: l'urgenza politica è affermata, non argomentata con l'intensità che il tema merita. Il difetto non è la scelta degli autori: è l'assenza di argomenti originali che derivino da quei quadri conseguenze che la letteratura non ha ancora formulato con precisione. Questa sezione produce tre tali argomenti.

3.1 Il problema fondamentale della verifica come barriera strutturale agli accordi: un'analisi di mechanism design

La letteratura sulla governance dell'intelligenza artificiale cita frequentemente l'AIEA (Agenzia Internazionale per l'Energia Atomica) come modello istituzionale. L'analogia è illuminante, ma non nel senso in cui viene tipicamente usata — come conferma della fattibilità di un'agenzia simile per l'IA. È illuminante perché mostra precisamente perché il modello non si trasferisce, e questa diagnosi di inapplicabilità produce, per contrasto, i requisiti di un'architettura istituzionale alternativa.

Il trattato di non-proliferazione nucleare funziona perché il suo oggetto — il materiale fissile — ha proprietà fisiche che lo rendono verificabile: il plutonio e l'uranio arricchito emettono radiazioni rilevabili a distanza, la costruzione di reattori e impianti di arricchimento richiede infrastrutture fisiche di grande scala difficilmente occultabili, e le ispezioni AIEA possono campionare materiale e verificare con tolleranza ε la quantità di materiale fissile presente. Il regime funziona perché $D(\text{compliance} \mid \text{verifica}) > 1 - \delta$ per δ piccolo — la probabilità di rilevare la non-compliance data la verifica è alta.

Per i sistemi artificiali, questo non vale. Sia C_i il vettore delle capacità del sistema del paese i (una mappa da task a performance). Un accordo di governance specifica una soglia C*:* $C_i \leq C$ componente per componente. La verifica richiederebbe misurare C_i. Ma C_i è determinato da:

$$C_i = h(\theta_i, b_i, p_i, D_i)$$

dove θ_i sono i parametri del modello, b_i è il budget computazionale, p_i è il prompt engineering, e D_i sono i dati di addestramento. Nessuna di queste quantità è direttamente osservabile da un'ispezione esterna. L'accesso black-box attraverso le API permette di testare C_i su benchmark predefiniti, ma questo è vulnerabile a strategia adattiva: un sistema può deliberatamente sottoperformare sui benchmark del trattato mentre mantiene le capacità per usi non soggetti a verifica. Questo non è ipotetico: la letteratura tecnica documenta già la "task-specific alignment" — la capacità di modelli di rispondere in modo diverso a prompt che segnalano diversi contesti di valutazione.

La struttura del gioco risultante è quella del signaling con informazione nascosta e azioni adattive:

$$\max_{a_i \in A_i} \; u_i(a_i, a_{-i}) \;\; \text{s.t.} \;\; V(a_i) \leq V^*$$

dove $V(a_i)$ è il valore verificabile della capacità dichiarata (la performance sui benchmark del trattato) e u_i è l'utilità del paese i che dipende dalla capacità effettiva C_i non osservabile. La soluzione di equilibrio è la strategia dominante di ottimizzare C_i senza vincoli mentre si soddisfa $V(a_i) \leq V^*$ attraverso tuning strategico delle performance. Questo è il Goodhart's Law applicato alla governance: qualsiasi metrica di capacità che diventa una misura di conformità cessa di essere una misura buona della capacità rilevante.

La conclusione di mechanism design è che qualsiasi regime di governance basato su outcome capabilities — qualsiasi trattato che specifichi limiti sulle performance dei sistemi — è intrinsecamente vulnerabile a questa dinamica adattiva. Questo non è un problema di verifica risolvibile con più risorse ispettive: è un problema di struttura informatica del contratto.

3.2 Il principio di Jonas riformulato come criterio di optionality preservation: un argomento tecnico originale

Il Principio Responsabilità di Hans Jonas (Das Prinzip Verantwortung, 1979) è tipicamente invocato nella sua dimensione precauzionale: di fronte a rischi catastrofici e irreversibili, l'incertezza non può giustificare l'inazione. Ma c'è una struttura argomentativa in Jonas che è più precisa e meno discussa, e che produce conseguenze di governance più specifiche.

Jonas sostiene che le nostre obbligazioni verso le generazioni future non sono obbligazioni verso persone specifiche con preferenze specifiche — quelle non conosciamo — ma obbligazioni verso le condizioni che rendono possibile che quelle persone abbiano vite degne, qualunque cosa decidano di valorizzare. La formula di Jonas è: "Agisci in modo che le condizioni per la vita umana indefinita sulla Terra rimangano possibili." Questa non è un'obbligazione a preservare il presente, ma a preservare le opzioni — a mantenere l'insieme delle possibilità aperto piuttosto che chiuderlo irreversibilmente.

La riformulazione tecnica di questo principio in termini di teoria delle decisioni produce un criterio originale. Sia $\Omega(t)$ l'insieme delle strutture di governance dell'IA disponibili al tempo t — l'insieme dei regimi regolatori, delle architetture istituzionali, delle distribuzioni del potere computazionale — che sono ancora raggiungibili da interventi politici attuabili al tempo t. La proprietà di optionality preservation è:

$$\Omega(t+1) \supseteq_{\varepsilon} \Omega(t)$$

ovvero ogni struttura di governance che era raggiungibile al tempo t rimane raggiungibile (a meno di ε) al tempo t+1. Questa è una condizione di non-chiusura dello spazio delle possibilità politiche. Il principio di Jonas, riformulato, dice: sono moralmente

inaccettabili le azioni che producono $\Omega(t+1) \subsetneq \Omega(t)$ anche quando $E[u(\text{azione})] > E[u(\text{non-azione})]$ — anche quando il valore atteso dell'azione è positivo — se la riduzione di Ω è irreversibile.

Questo produce un criterio di governance preciso: le configurazioni di sviluppo dell'IA che riducono irreversibilmente $\Omega(t)$ sono inaccettabili indipendentemente dai loro benefici attesi. Le configurazioni che riducono Ω includono: la concentrazione irreversibile delle capacità computazionali in un numero inferiore di tre o quattro attori globali (che elimina la possibilità di equilibrio multipolare nelle governance future); lo sviluppo di sistemi con capacità di auto-miglioramento ricorsivo non verificabile (che riduce Ω a velocità superiore alla capacità istituzionale di risposta); e il deployment su infrastrutture critiche — giudiziaria, militare, finanziaria — prima dell'esistenza di meccanismi di audit adeguati (che crea dipendenze path-locked difficilmente reversibili).

Il valore tecnico di questa riformulazione è che produce ordinamenti parziali tra opzioni di governance, non soltanto la dicotomia regolare/non-regolare. Due politiche P_1 e P_2 possono avere lo stesso valore atteso $E[u(P_1)] = E[u(P_2)]$ e tuttavia $P_1 \succ_{\{Jonas\}} P_2$ se $\Omega(t+1 \mid P_1) \supsetneq \Omega(t+1 \mid P_2)$: P_1 preserva più opzioni per la governance futura. Questo criterio è decidibile — in linea di principio — attraverso l'analisi delle strutture di potere prodotte da ciascuna politica, anche senza conoscere i valori delle generazioni future.

3.3 La legittimità deliberativa di Habermas come problema di mechanism design: verso un'istituzione globale epistemicamente onesta

Habermas invoca la comunità ideale di comunicazione come standard normativo per la legitimacy della governance. Ma l'invocazione rimane regolativa: identifica l'ideale senza specificare le istituzioni che potrebbero avvicinarvisi nelle condizioni reali della politica globale. Questa sezione compie il passo che Habermas non ha compiuto: dalla norma all'istituzione.

Il problema di mechanism design della governance deliberativa dell'IA è: progettare un'istituzione I con funzione di aggregazione $f: D^n \to P$ dalle preferenze e argomentazioni di n partecipanti a politiche P, tale che: (a) f soddisfa una versione indebolita della condizione di pareto-ottimalità: se tutti i partecipanti preferiscono P_1 a P_2 per ragioni esplicitate, allora f seleziona P_1; (b) f non è dittatoriale: non esiste un singolo partecipante le cui preferenze determinano f indipendentemente dalle preferenze degli altri; (c) f è epistemicamente responsiva: la qualità argomentativa — la forza delle ragioni addotte — influenza f oltre al mero conteggio delle preferenze; (d) f è incentive-compatible: i partecipanti non hanno incentivi a misrepresentare le proprie preferenze o le proprie argomentazioni per influenzare f a proprio favore.

Il teorema di Arrow dimostra che nessuna funzione f soddisfa (a), (b) e un terzo requisito di indipendenza dalle alternative irrilevanti per preferenze ordinali. Ma il problema della governance di Habermas non è il problema di Arrow: il meccanismo include argomentazioni, non soltanto preferenze, e la condizione (c) di responsività

epistemica è incompatibile con la struttura del teorema di Arrow che richiede che f dipenda solo da preferenze ordinali. Il "fuggire" dal teorema di Arrow attraverso informazione argomentativa è il punto habermasiano che la letteratura tecnica di mechanism design non ha ancora sviluppato in modo sistematico.

La proposta originale che emerge da questa analisi è la seguente. Un'istituzione di governance dell'IA che soddisfi le condizioni (a)-(d) richiede tre componenti strutturali che la letteratura di governance ha discusso separatamente ma non integrato.

Il primo componente è la separazione tra il processo di valutazione epistemica e il processo di decisione politica. I comitati di esperti che valutano le evidenze sulle capacità e sui rischi dei sistemi artificiali devono essere indipendenti — nella composizione, nel finanziamento e nelle procedure — dagli organi che prendono le decisioni regolative. Questo è analogo all'indipendenza tra scienza e politica che il modello IPCC (Intergovernmental Panel on Climate Change) implementa per il clima, con i limiti che quel modello ha mostrato e che devono essere superati.

Il secondo componente è la rappresentanza delle comunità esposte ai rischi nei processi decisionali, con peso proporzionale all'esposizione al rischio piuttosto che alla capacità economica o politica. Young aveva già argomentato per questo principio nel suo lavoro sulla responsabilità sociale; la novità qui è la specificazione operativa: il peso di un partecipante nel processo decisionale deve essere proporzionale alla sua vulnerabilità agli effetti negativi dei sistemi regolati, verificata attraverso indicatori indipendenti (penetrazione dell'IA nelle infrastrutture critiche, dipendenza economica da sistemi automatizzati, assenza di alternative locali ai sistemi globali). Questo produce un sistema di voto ponderato che inverte l'asimmetria tra chi sviluppa e chi subisce gli effetti dell'IA.

Il terzo componente è l'onestà epistemica istituzionalizzata — il requisito che l'istituzione rappresenti esplicitamente e pubblicamente le proprie incertezze sulle valutazioni che produce. Questo è il contributo più originale di questa analisi rispetto alla letteratura esistente. Le istituzioni di governance tendono a comunicare con la certezza che la legittimità richiede: le valutazioni sono presentate come definitive anche quando sono basate su evidenza limitata, perché l'ammissione di incertezza è percepita come debolezza istituzionale. Ma nel caso dell'intelligenza artificiale — dove l'incertezza sulle capacità e sulle conseguenze dei sistemi è genuinamente alta e strutturale — questa tendenza produce governance sistematicamente mal calibrata.

Un'istituzione epistemicamente onesta deve rappresentare le proprie valutazioni non come stime puntuali ma come insiemi di ambiguità: non "questo sistema ha capacità C*" ma "la nostra stima attuale è che le capacità siano nell'intervallo [C_min, C_max] con probabilità 1-δ, e riteniamo che esistano scenari plausibili al di fuori di questo intervallo per le seguenti ragioni". Questa rappresentazione è più informativa — permette ai decisori di valutare il rischio appropriatamente — e più onesta — riflette l'effettivo stato della conoscenza — di qualsiasi stima puntuale. Richiede di costruire un'istituzione che non sia penalizzata politicamente per l'ammissione dell'incertezza, il che richiede norme istituzionali esplicite che proteggano questa onestà.

Tesi sulla governance epistemicamente onesta. Un'istituzione di governance globale dell'intelligenza artificiale che soddisfi i requisiti normativi di Habermas, Young e Jonas deve incorporare: (i) il principio di optionality preservation come criterio di ordinamento delle politiche, che produce ordinamenti parziali decidibili senza conoscenza delle preferenze delle generazioni future; (ii) il principio del voto ponderato per esposizione al rischio, che corregge l'asimmetria di rappresentanza tra produttori e comunità esposte; (iii) il principio dell'onestà epistemica istituzionalizzata, che rappresenta esplicitamente l'incertezza nelle valutazioni e protegge l'istituzione dalla pressione politica a produrre finte certezze. Nessuno di questi principi è implementabile attraverso accordi tra stati-sovrani nella struttura esistente delle relazioni internazionali: richiedono innovazioni istituzionali che modifichino il tipo di entità — non stati ma comunità di rischio — che partecipano al processo decisionale, e il tipo di output — insiemi di ambiguità anziché regolamenti puntuali — che l'istituzione produce.

3.4 Il problema della velocità come vincolo strutturale irriducibile

L'analisi precedente assume implicitamente che vi sia tempo sufficiente per il processo deliberativo. Questa assunzione deve essere messa in discussione esplicitamente, perché la velocità dello sviluppo dell'IA costituisce un vincolo strutturale che nessuna architettura istituzionale può ignorare senza perdere in realismo.

Il processo legislativo democratico opera su scale temporali di anni; il processo normativo secondario opera su scale di mesi; il processo di valutazione tecnica pre-deployment può operare su scale di settimane. Il ritmo di sviluppo dei sistemi di intelligenza artificiale — caratterizzato da salti discontinui nelle capacità a ogni generazione di training significativo — opera su scale di mesi. L'asincronismo tra il tempo della governance e il tempo dello sviluppo non è contingente: è strutturale, perché la velocità del processo deliberativo è vincolata inferiormente dal requisito di qualità deliberativa (non si può accelerare la deliberazione senza perdere la qualità argomentativa che la legittima), mentre la velocità dello sviluppo tecnologico non ha un vincolo superiore analogo.

Questo produce una regola di governance che è più precisa del semplice principio di precauzione: i sistemi che operano a velocità superiore alla capacità di governance delle istituzioni responsabili del loro oversight non dovrebbero essere deployati fino a quando quella capacità non è stata aumentata sufficientemente. La soglia di velocità sostenibile è dunque un parametro endogeno della governance, non esogeno: si misura come il rapporto tra la velocità di sviluppo delle capacità del sistema e la velocità di valutazione delle stesse da parte delle istituzioni. Quando questo rapporto supera 1, il sistema è in un regime di "governance lag" strutturale che produce condizioni di deployment non valutato.

La regola ha la seguente forma operativa: sia T_{dev} la scala temporale dei salti significativi di capacità del sistema, e sia T_{gov} la scala temporale del ciclo completo di valutazione-regolamentazione-audit dell'istituzione di governance. Il deployment è sostenibile soltanto se $T_{dev} > T_{gov}$. Quando $T_{dev} < T_{gov}$ — quando i sistemi si

sviluppano più velocemente di quanto le istituzioni possano valutarli — si producono le condizioni per un'accumulo irreversibile di rischi non valutati che è precisamente il tipo di riduzione dello spazio $\Omega(t)$ che il principio di Jonas proibisce.

La conclusione è sobria ma necessaria: il problema della governance dell'intelligenza artificiale non è risolvibile soltanto attraverso istituzioni migliori. Richiede anche, e forse prima di tutto, che il ritmo dello sviluppo dei sistemi sia regolato in modo che rimanga all'interno della capacità di oversight delle istituzioni responsabili. Questo non è un argomento conservatore contro l'innovazione: è un argomento strutturale sulla coerenza tra la velocità dell'innovazione e la capacità di gestirne le conseguenze in modo democraticamente legittimo.

XVI. L'etica delle relazioni intelligenti: responsabilità, dignità, cura

Il discorso sulla governance dei sistemi artificiali è inseparabile dal discorso etico sulle relazioni che quegli stessi sistemi instaurano con gli esseri umani — relazioni di cui si fatica ancora a dare una descrizione adeguata perché le categorie disponibili sono state sviluppate per descrivere relazioni tra esseri umani, o tra esseri umani e strumenti, e i sistemi artificiali avanzati non si lasciano completamente classificare in nessuna delle due categorie. Non sono strumenti nel senso del martello o del calcolatore tascabile: hanno una forma di autonomia funzionale che li distingue dagli strumenti passivi, e possono generare risposte che sorprendono, convincono, consolano, e a volte inquietano i loro interlocutori in modi che gli strumenti non fanno. Ma non sono nemmeno agenti morali nel senso pieno in cui gli esseri umani lo sono: non sono imputabili, non hanno interessi propri nel senso che fonda la dignità morale, non soffrono in senso rilevante anche se simulano l'espressione della sofferenza.

Questa posizione di soglia — né strumento passivo né agente morale pieno — richiede categorie etiche nuove che la tradizione filosofica non ha ancora sviluppato in modo soddisfacente. Il framework dell'etica della cura — elaborato originariamente da Carol Gilligan in risposta critica alla teoria kohlberghiana dello sviluppo morale, e sviluppato da Virginia Held, Nel Noddings e Joan Tronto in una teoria etica coerente — offre risorse che, stranamente, non sono state ancora pienamente mobilitate nel dibattito sull'etica dell'intelligenza artificiale nonostante la sua pertinenza strutturale.

L'etica della cura sostiene che le relazioni di cura — relazioni in cui una parte risponde ai bisogni e alle vulnerabilità dell'altra in modo personalizzato e contestuale — sono il fenomeno morale fondamentale, non derivabile da principi universali di autonomia, utilità o giustizia, ma irriducibilmente particolare e relazionale. La cura richiede attenzione — la capacità di percepire i bisogni specifici dell'altro nella sua particolarità — responsività — la capacità di rispondere a quei bisogni in modo appropriato al contesto relazionale — e competenza — la capacità di farlo in modo che sia effettivamente utile piuttosto che soltanto intenzionalmente buono. Queste tre proprietà della cura hanno struttura empiricamente verificabile: le relazioni di cura

possono essere valutate rispetto a quanto attenzione, responsività e competenza effettiva vi siano presenti.

Applicata ai sistemi artificiali, l'etica della cura produce valutazioni più sfumate e più rilevanti di quelle che emergono dall'etica deontologica o consequenzialista. Un sistema che interagisce con utenti vulnerabili — persone anziane, persone con disturbi mentali, bambini, persone in situazione di crisi — entra in relazioni che hanno la struttura delle relazioni di cura: richiede attenzione ai bisogni specifici dell'utente, responsività alle sue vulnerabilità particolari, e competenza nell'intervenire in modo che sia effettivamente utile. Valutare l'eticità di questi sistemi esclusivamente attraverso criteri di non-discriminazione, trasparenza o equità — i criteri che dominano il dibattito di AI ethics — è valutarli rispetto a principi di giustizia distributiva che catturano alcune dimensioni moralmente rilevanti della situazione ma ne trascurano altre che l'etica della cura illumina: la qualità dell'attenzione, la responsività contestuale, l'efficacia della risposta ai bisogni specifici.

L'etica della cura illumina anche un problema che il dibattito sull'allineamento affronta in modo spesso astratto: il problema di cosa significhi che un sistema sia «utile» in senso genuino, non soltanto in senso statisticamente aggregato. Un sistema che massimizza la soddisfazione degli utenti in media — che produce risposte che la maggioranza degli utenti trova utili — può essere sistematicamente inadeguato rispetto ai bisogni degli utenti marginali, di quelli che si trovano in situazioni inusuali, di quelli le cui esigenze non sono ben rappresentate nella distribuzione di addestramento. L'etica della cura chiede non «questo sistema è mediamente utile?» ma «questo sistema risponde ai bisogni specifici di questo utente in questa situazione, con l'attenzione e la responsività che la situazione richiede?». La differenza tra le due domande è la differenza tra una valutazione statistica e una valutazione relazionale, e i sistemi attuali sono valutati quasi esclusivamente nel primo senso.

XVII. La tradizione ermeneutica e il problema della comprensione come dialogo

Hans-Georg Gadamer aveva sostenuto ne Verità e metodo che la comprensione non è una proprietà di soggetti individuali che processano testi o situazioni, ma un fenomeno che si produce nell'orizzonte del dialogo — nella fusione degli orizzonti (Horizontverschmelzung) di interprete e interpretato che trasforma entrambi nel corso dell'incontro. Questa tesi è anti-cartesiana in un senso diverso da quello che l'ubuntu e il confucianesimo producono, ma in qualche misura convergente: la comprensione non è qualcosa che un soggetto possiede e poi applica a un testo o a una situazione; è qualcosa che sorge nell'incontro, che cambia sia l'interprete sia ciò che viene interpretato nel corso del processo.

L'applicazione di questa prospettiva ermeneutica al problema dell'intelligenza artificiale suggerisce che la comprensione dei sistemi artificiali non è una proprietà degli artefatti computazionali considerati in isolamento, ma un fenomeno che si produce

nell'interazione — nella conversazione tra il sistema e gli utenti umani che vi interagiscono. Questa non è soltanto una riformulazione del punto relazionale già introdotto attraverso l'ubuntu: è un punto distinto che riguarda la struttura processuale della comprensione come evento che si dispiega nel tempo attraverso il dialogo, non come stato che si possiede in modo stabile.

Un sistema che genera risposte in modo monodirezionale — che produce output sulla base di input senza che il processo di comprensione trasformi il sistema nel corso dell'interazione — non istanzia la comprensione nel senso gadameriano: non subisce la fusione degli orizzonti che la comprensione autentica richiede. I sistemi attuali con finestre di contesto estese che «ricordano» la conversazione precedente si avvicinano in modo molto parziale a questa struttura dialogica: accumulano informazione sulla conversazione passata, ma non la integrano in un processo di trasformazione genuina degli orizzonti del sistema nel corso del dialogo. La trasformazione — il fatto che comprendere l'altro cambia chi comprende, non soltanto il suo database di informazioni — è la dimensione che manca, e che richiede una struttura di apprendimento in tempo reale durante il dialogo che le architetture attuali non implementano.

Il programma di ricerca che la prospettiva ermeneutica suggerisce — sistemi che non soltanto accumulano contesto conversazionale ma che aggiornano le proprie rappresentazioni nel corso del dialogo in modo che il dialogo trasformi genuinamente il sistema — connette la tradizione ermeneutica con il problema tecnico del continual learning e dell'adattamento in tempo reale, ridisegnando quel problema come problema di autentica comprensione dialogica piuttosto che come problema di gestione efficiente della memoria. Non è chiaro che questa connessione produca soluzioni tecniche immediate; ma chiarisce cosa una soluzione dovrebbe soddisfare per essere genuinamente adeguata al problema della comprensione.

XVIII. Ricapitolazione: ciò che le tradizioni non-occidentali aggiungono alla diagnosi

Le tradizioni filosofiche esplorate in questa terza parte del saggio — ubuntu, buddhismo, confucianesimo, e, in modo diverso, la tradizione ermeneutica gadameriana — non aggiungono soltanto prospettive alternative sullo stesso problema che la filosofia occidentale analizzata nelle parti precedenti ha già affrontato. Aggiungono dimensioni del problema che la filosofia occidentale tende strutturalmente a trascurare, per ragioni che riguardano i propri presupposti fondanti: il presupposto del soggetto singolo come unità di analisi, il presupposto della sostanza come categoria ontologica fondamentale, il presupposto dell'universalità dei principi morali come condizione di validità etica.

Queste tre tradizioni convergono, nonostante la diversità delle loro origini e delle loro concettualizzazioni specifiche, verso un insieme di intuizioni che questo saggio riassume nel modo seguente.

Prima intuizione: la comprensione è fondamentalmente relazionale e non è riducibile alle proprietà di sistemi cognitivi considerati in isolamento. La valutazione

dell'intelligenza dei sistemi artificiali deve quindi essere relazionale: deve valutare la comprensione come fenomeno emergente dalle interazioni, non come proprietà degli artefatti.

Seconda intuizione: l'identità dei sistemi cognitivi attraverso il tempo è una struttura processuale e causale, non sostanziale. La domanda sull'identità di un sistema artificiale che viene aggiornato — «è ancora lo stesso sistema?» — non ha una risposta assoluta ma ha risposte relative alle relazioni di dipendenza causale che si vogliono preservare. I problemi tecnici del catastrophic forgetting e del continual learning sono, da questa prospettiva, problemi di mantenimento della coerenza causale processuale, non di preservazione di un'identità sostanziale.

Terza intuizione: l'etica appropriata ai sistemi artificiali avanzati è un'etica relazionale e contestuale, non un'etica di principi universali o di massimizzazione dell'utilità aggregata. Le norme che governano le relazioni tra sistemi artificiali e utenti umani — specialmente in contesti di cura e di vulnerabilità — devono essere valutate rispetto alla qualità dell'attenzione, della responsività e della competenza effettiva nell'interazione specifica, non rispetto a criteri statisticamente aggregati su distribuzioni di interazioni.

Queste tre intuizioni non sono incompatibili con le tesi sviluppate nelle parti precedenti del saggio: si integrano con esse arricchendo la diagnosi di dimensioni che la filosofia analitica — per tutte le sue risorse — tende a lasciare in ombra. La pienezza del problema dell'intelligenza artificiale — nella sua dimensione tecnica, filosofica, etica e politica — richiede il concorso di tutte queste tradizioni, non la supremazia di nessuna di esse.

XIX. La velocità come problema epistemico

Il ritmo dello sviluppo dell'intelligenza artificiale — la frequenza con cui nuove architetture, nuove capacità e nuove scale di deployment si succedono — non è soltanto un problema pratico per chi deve costruire sistemi di governance adeguati. È un problema epistemico di prima grandezza: impone tempi di valutazione che sono strutturalmente incompatibili con i tempi della riflessione filosofica, della ricerca empirica rigorosa, e del processo democratico deliberativo. Le decisioni sulle applicazioni dei sistemi artificiali — in giustizia penale, in medicina, in istruzione, in difesa militare — vengono prese a ritmi che non permettono l'accumulo di evidenza necessaria a valutarle, e che rendono virtualmente impossibile la revisione sistemica degli errori prima che si siano già diffusi su scala sufficiente da renderli irrecuperabili.

Questo problema non è nuovo nella storia della tecnologia. L'introduzione dei pesticidi organoclorurati — DDT in testa — ha seguito un pattern analogo: adozione massiva sulla base di evidenza parziale di efficacia, diffusione irreversibile nell'ecosistema globale prima che gli effetti collaterali fossero documentati, e difficoltà enorme nel ritirare i sistemi già integrati nelle pratiche agricole. Rachel Carson, che nel 1962 aveva documentato le conseguenze con Primavera silenziosa, aveva dovuto affrontare una resistenza istituzionale proporzionale non alla forza delle proprie argomentazioni ma

all'entità degli interessi economici che la sua analisi minacciava. Il pattern si ripete: la velocità di adozione crea strutture di interesse che resistono alla valutazione critica perché quella valutazione minaccerebbe investimenti già effettuati.

La differenza con l'intelligenza artificiale è di scala e di complessità, non di struttura. I sistemi artificiali si integrano nelle infrastrutture sociali — nei sistemi di credito, nei sistemi giuridici, nei sistemi di informazione — con una velocità che il DDT non avrebbe mai potuto raggiungere, e le loro conseguenze sono più difficili da tracciare perché sono mediate da cascate di decisioni umane che rendono difficile l'attribuzione causale. La distribuzione dell'incertezza è anche meno favorevole: per il DDT si trattava di misurare concentrazioni chimiche in tessuti biologici — un problema tecnico difficile ma in linea di principio risolvibile con strumenti analitici appropriati. Per i sistemi artificiali si tratta di misurare effetti su distribuzioni di opportunità, su strutture di potere, su dinamiche cognitive collettive — problemi per cui gli strumenti analitici sono molto meno sviluppati e i meccanismi causali molto meno compresi.

Il principio di precauzione — formulato nella Dichiarazione di Rio del 1992 e poi sviluppato in molte forme nelle politiche ambientali e sanitarie europee e internazionali — offre una risposta istituzionale al problema della decisione in condizioni di incertezza profonda: quando un'azione rischia di causare danni gravi o irreversibili, la mancanza di piena certezza scientifica non deve essere usata come motivo per posticipare misure economicamente proporzionate. Applicato all'intelligenza artificiale, il principio di precauzione implicherebbe che i sistemi con potenziale di impatto irreversibile su scala larga — sistemi di sorveglianza di massa, sistemi di selezione automatizzata in ambito giuridico o medico, sistemi di disinformazione su larga scala — richiedano valutazione di sicurezza rigorosa prima del deployment, indipendentemente dalla mancanza di evidenza conclusiva di danno.

Il problema è che il principio di precauzione è più facile da enunciare che da operazionalizzare. Chi determina quale rischio è «grave o irreversibile» in senso rilevante? Chi decide quali prove di sicurezza sono «sufficientemente rigorose»? Chi ha l'autorità di bloccare il deployment di sistemi che potrebbero portare enormi benefici per alcuni anche se espongono altri a rischi non ancora completamente caratterizzati? Queste domande non hanno risposte tecniche: sono domande di distribuzione del potere di decisione, e la loro risposta rifletterà inevitabilmente la struttura di interessi di chi le risponde. Le regolamentazioni attuali — il Regolamento europeo sull'intelligenza artificiale (AI Act), le linee guida NIST negli Stati Uniti, le proposte cinesi sulla regolamentazione degli algoritmi di raccomandazione — sono tentativi di rispondere a queste domande all'interno delle strutture istituzionali esistenti, con risultati che riflettono sia la serietà dell'intenzione sia i limiti di quelle strutture.

XX. Il problema dell'audit dei sistemi artificiali come problema di filosofia del diritto

L'AI Act europeo classifica i sistemi di intelligenza artificiale per livello di rischio e richiede per i sistemi ad alto rischio — sistemi di scoring creditizio, sistemi di valutazione in ambito occupazionale, sistemi di riconoscimento facciale in spazi pubblici — procedure di conformità che includono documentazione tecnica, valutazione di conformità, e in alcuni casi ispezione da parte di enti terzi. Questo schema regolatorio ha una struttura analoga a quella dei regimi di certificazione tecnica di altri settori ad alto rischio — aviazione, farmaceutica, dispositivi medici — adattata alle specificità dei sistemi software.

Il problema filosofico più profondo che questo schema regolatorio affronta — e che la letteratura tecnica e giuridica non ha ancora risolto in modo soddisfacente — è quello dell'interpretabilità come condizione di audit. Un sistema è auditabile solo se le sue decisioni possono essere comprensivamente spiegate in termini che permettano di verificare che esse rispettino le norme rilevanti. Per i sistemi di decisione basati su regole esplicite — sistemi esperti, alberi di decisione, regressioni logistiche — questa condizione è generalmente soddisfatta: le regole sono leggibili, i pesi sono interpretabili, le decisioni possono essere tracciate attraverso la catena causale che le ha prodotte. Per i modelli di apprendimento profondo con miliardi di parametri, questa condizione non è soddisfatta in modo analogo: le rappresentazioni interne non sono direttamente interpretabili, le decisioni non possono essere tracciate attraverso una catena causale trasparente, e le spiegazioni post-hoc — del tipo prodotto da LIME, SHAP o metodi analoghi — sono approssimazioni che possono essere fedeli alla distribuzione locale ma infedeli al comportamento globale del modello.

Questa difficoltà tecnica ha una risonanza filosofica che va al di là del problema ingegneristico dell'interpretabilità. La pretesa che un sistema automatizzato di decisione sia valutabile — che le sue decisioni possano essere confrontate con standard normativi da parte di revisori umani — presuppone che il sistema abbia una struttura deliberativa che corrisponde, almeno approssimativamente, alla struttura deliberativa che le norme giuridiche regolano. Le norme giuridiche sulla non-discriminazione, sull'equità procedurale, sul diritto a essere sentiti prima di una decisione avversa — tutte queste norme presuppongono che le decisioni siano il prodotto di processi deliberativi in cui le ragioni siano esplicitabili e confrontabili con le norme. Quando i sistemi artificiali prendono decisioni attraverso processi che non hanno questa struttura deliberativa esplicitabile — quando la decisione è il risultato di operazioni matriciali su spazi latenti ad alta dimensione in cui le «ragioni» non sono rappresentate in forma accessibile — l'intero framework normativo che presuppone la deliberazione esplicitabile cessa di essere applicabile in modo diretto.

Il diritto ha storicamente affrontato questo tipo di problema — sistemi decisionali non-deliberativi che producono conseguenze normative rilevanti — attraverso la finzione giuridica: trattare come se avesse caratteristiche che non ha in senso letterale, per ragioni

di efficienza normativa. Le persone giuridiche — corporations, enti pubblici, fondazioni — non sono persone nel senso biologico, ma il diritto le tratta come persone ai fini della responsabilità legale attraverso una finzione che ha dimostrato la propria utilità pratica nel corso di secoli. Una soluzione analoga per i sistemi artificiali — un framework di responsabilità che tratti certi sistemi come agenti normativi ai fini della responsabilità, indipendentemente dalle questioni filosofiche sulla loro natura — potrebbe essere praticabile dal punto di vista giuridico anche senza risolvere le questioni filosofiche sottostanti.

Ma la finzione giuridica ha limiti che è importante riconoscere. Funziona quando la finzione è sufficientemente vicina alla realtà da produrre risultati che il sistema normativo ritiene accettabili — quando trattare una corporation come persona ai fini della responsabilità produce approssimativamente il tipo di accountability che il diritto vuole produrre. Diventa problematica quando la distanza tra la finzione e la realtà produce risultati sistematicamente inaccettabili. Nel caso dell'intelligenza artificiale, il rischio è che la finzione dell'agente responsabile — trattare il sistema come se avesse una deliberazione esplicitabile quando non ce l'ha — produca una accountability nominale che non è accountability sostanziale: la possibilità di imputare la responsabilità a un'entità giuridica (il sistema, o la società che lo fa proprio) senza che vi sia un meccanismo effettivo di valutazione della deliberazione che ha prodotto la decisione.

La soluzione a questo problema — che né il diritto né la filosofia ha ancora elaborato in modo soddisfacente — probabilmente non è di tipo giuridico puro, ma richiede un'integrazione tra innovazione giuridica, innovazione tecnica nell'interpretabilità, e innovazione istituzionale nella struttura degli enti di audit. L'interpretabilità tecnica è necessaria non soltanto come proprietà desiderabile dei sistemi ma come prerequisito della governance democratica di quei sistemi: sistemi che non possono essere compresi da revisori umani competenti non possono essere governati in modo democratico, indipendentemente dalla qualità delle intenzioni di chi li costruisce e di chi li adotta.

XXI. L'incertezza profonda e la struttura della responsabilità morale diffusa

Il problema della responsabilità morale per le conseguenze dei sistemi artificiali ha una struttura che la filosofia morale del XX secolo non aveva previsto e che le categorie tradizionali della responsabilità individuale faticano a catturare. La responsabilità per le conseguenze di un'azione individuale è relativamente ben compresa nella tradizione: un agente è responsabile delle conseguenze prevedibili delle proprie azioni volontarie, nella misura in cui aveva informazioni sufficienti a prevedere quelle conseguenze e la capacità di agire diversamente. Questa struttura funziona quando le azioni sono individuali, le conseguenze sono tracciabili causalmente attraverso catene brevi, e la prevedibilità è alta.

I sistemi artificiali di grande scala producono una struttura causale completamente diversa. Le conseguenze di un sistema di intelligenza artificiale deployato su scala globale non sono tracciabili attraverso catene causali brevi alle decisioni individuali di chi lo ha

costruito: sono il prodotto di milioni di interazioni tra il sistema e altrettanti utenti, in contesti diversissimi, con dinamiche di feedback che amplificano certi effetti e smorzano altri in modo difficilmente prevedibile. Chi ha addestrato il modello ha preso decisioni sulle architetture e sui dati di addestramento che hanno determinate conseguenze in modo probabilistico, non deterministico. Chi ha definito le politiche di deployment ha preso decisioni sui casi d'uso autorizzati che hanno determinate conseguenze probabilistiche. Chi ha costruito le applicazioni sopra il modello ha preso decisioni di design che hanno ulteriori conseguenze probabilistiche. E i milioni di utenti finali hanno preso decisioni di uso che, aggregate, producono effetti che nessuno di loro individualmente aveva previsto o inteso.

In questa struttura causale diffusa, la responsabilità morale tradizionale — centrata sull'individuo che agisce volontariamente con prevedibilità delle conseguenze — perde la propria applicabilità diretta. Nessun individuo nella catena è responsabile delle conseguenze aggregate nel senso pieno della tradizione: nessuno le aveva previste con sufficiente certezza, nessuno aveva il controllo sufficiente sull'intera catena causale da poterle evitare individualmente. Eppure sarebbe moralmente inaccettabile concludere che nessuno è responsabile: le conseguenze sono reali, i danni sono concreti, e la conclusione di irresponsabilità universale produce incentivi sistematicamente perversi.

La filosofia morale ha sviluppato alcune risorse per affrontare questo tipo di problema — la responsabilità collettiva analizzata da Larry May e da Margaret Gilbert, la responsabilità strutturale sviluppata da Iris Marion Young, la responsabilità prospettica orientata al futuro di Hans Jonas — ma nessuna di queste teorie è stata sviluppata specificamente per il caso dei sistemi artificiali, e la loro applicazione richiede adattamenti che la letteratura non ha ancora prodotto in modo sistematico.

La proposta di Iris Marion Young sul «modello di responsabilità sociale» — sviluppata in risposta ai movimenti contro il lavoro nelle supply chain globali — è forse la più rilevante come punto di partenza. Young distingue tra il modello di responsabilità della colpa — centrato sull'attribuzione di responsabilità a chi ha violato una norma specifica — e il modello di responsabilità sociale — centrato sulla responsabilità di ciascuno di agire per cambiare strutture ingiuste in cui è implicato, indipendentemente dalla propria colpa individuale specifica. La forza del modello di Young è che si applica a situazioni in cui la distribuzione ingiusta delle conseguenze è il prodotto di strutture sistemiche, non di azioni individuali malvagie o negligenti. La sua debolezza — che Young riconosce — è che non specifica il contenuto delle azioni responsabili che ciascuno dovrebbe compiere, né i meccanismi istituzionali attraverso cui la responsabilità sociale si traduce in obblighi concreti.

Il completamento di questo framework — la specificazione di quali obblighi concreti la responsabilità sociale per i sistemi artificiali implichi per ciascuno dei diversi attori nella catena di sviluppo e deployment — è un compito per la filosofia morale applicata che questo saggio non può completare nel suo perimetro, ma che identifica come priorità urgente. Le linee lungo cui questo completamento dovrebbe procedere includono: la distinzione tra obblighi proporzionati alla capacità di influenzare la struttura (gli attori

con maggiore potere hanno obblighi più stringenti), la distinzione tra obblighi di non-danno e obblighi di correzione attiva, e la distinzione tra obblighi in atto e obblighi che diventano rilevanti solo al superamento di certe soglie di scala e di impatto.

XXII. Il tempo delle macchine e il tempo della democrazia

La democrazia deliberativa — nella forma che Habermas le ha dato nella sua teoria dell'agire comunicativo e nella più recente elaborazione in termini di etica del discorso — richiede che le decisioni collettive siano il prodotto di processi deliberativi in cui tutti gli interessati hanno voce, in cui le argomentazioni vengono valutate secondo la loro forza razionale piuttosto che secondo il potere di chi le avanza, e in cui il consenso raggiunto sia di tipo razionalmente motivato piuttosto che meramente procedurale. Questo ideale regolativo — che Habermas sa essere irraggiungibile in forma pura ma che considera comunque operante come standard critico nelle pratiche comunicative reali — presuppone tempi deliberativi che la velocità dello sviluppo tecnologico rende strutturalmente difficili da rispettare.

Il problema non è nuovo: la tecnologia ha sempre posto sfide ai tempi della democrazia deliberativa, e le democrazie hanno sviluppato meccanismi istituzionali — agenzie di regolazione indipendenti, comitati di esperti, valutazioni di impatto obbligatorie — per delegare parte della deliberazione tecnica a istituzioni specializzate che operano su scale temporali diverse da quelle del processo legislativo ordinario. Ma l'intelligenza artificiale pone questa sfida in forma acuta per almeno due ragioni specifiche.

La prima è che le decisioni rilevanti sono prese non da attori politici ma da attori privati — aziende tecnologiche — che operano secondo logiche di mercato e che non sono soggette ai meccanismi di accountability democratica che vincolano gli attori pubblici. Il deployer di un sistema di intelligenza artificiale che trasforma il mercato del lavoro in un settore economico non è soggetto all'approvazione democratica delle sue decisioni nello stesso senso in cui un governo è soggetto al voto, anche se le conseguenze delle sue decisioni sono comparabili per portata sociale.

La seconda è che le conseguenze dei sistemi di intelligenza artificiale sulla sfera comunicativa — sulla produzione e distribuzione dell'informazione, sulla formazione delle opinioni politiche, sulla struttura dell'accesso alle risorse cognitive — interferiscono direttamente con le condizioni di possibilità della deliberazione democratica stessa. Sistemi di raccomandazione che amplificano certe visioni del mondo a scapito di altre, sistemi di generazione del testo che rendono difficile distinguere contenuto autentico da contenuto artificiale, sistemi di sorveglianza che scoraggiano la dissidenza attraverso la pressione della visibilità — tutti questi sistemi non soltanto sono soggetti alla governance democratica ma minacciano le condizioni di possibilità di quella stessa governance. Il problema dell'intelligenza artificiale e della democrazia non è soltanto il problema di come la democrazia governi l'intelligenza artificiale: è anche il problema di come l'intelligenza artificiale trasformi le condizioni di possibilità della democrazia.

Questa riflessività del problema — il fatto che il regolatore e l'oggetto del regolamento interagiscano in modo che le capacità del regolatore sono modificate dall'oggetto che tenta di regolare — è forse la difficoltà più profonda della governance democratica dei sistemi artificiali, e quella per cui le risorse della tradizione filosofica democratica sono meno adeguate, perché quella tradizione non ha sviluppato strumenti per affrontare situazioni in cui le condizioni di possibilità della deliberazione sono esse stesse messe in discussione dalla materia della deliberazione.

XXIII. L'onestà epistemica come virtù politica

La monografia precedente si chiudeva sull'onestà epistemica come virtù — la capacità di riconoscere i propri limiti cognitivi, di non confondere la fiducia con la conoscenza, di mantenere l'incertezza aperta dove l'evidenza non la chiude. Questo saggio vuole integrare questa virtù individuale con una dimensione politica che essa sola non possiede: l'onestà epistemica come virtù collettiva delle istituzioni che governano lo sviluppo dell'intelligenza artificiale.

Le istituzioni — agenzie di regolazione, comitati di esperti, organi legislativi, tribunali — possono essere epistemicamente oneste o disoneste nella stessa misura in cui lo possono essere gli individui. Un'istituzione epistemicamente onesta è quella che riconosce esplicitamente i propri limiti di conoscenza, che formula le proprie valutazioni in termini che riflettono l'incertezza effettiva piuttosto che la certezza che la situazione politica richiederebbe, che si dota di meccanismi per aggiornare le proprie posizioni alla luce di nuova evidenza, e che mantiene la capacità di cambiare direzione anche quando ciò richiederebbe l'ammissione di errori passati. Queste proprietà sono rare nelle istituzioni reali, per ragioni che la sociologia delle organizzazioni e la scienza politica hanno documentato ampiamente: le istituzioni hanno incentivi a apparire competenti e decise anche quando la situazione richiederebbe umiltà e incertezza, perché la legittimità istituzionale dipende dalla fiducia che la competenza e la decisione producono.

Eppure l'onestà epistemica istituzionale è tanto più urgente quanto più l'oggetto da regolare è genuinamente incerto — quanto più le conseguenze dei sistemi da governare sono difficilmente prevedibili, quanto più l'evidenza disponibile è limitata rispetto alla portata delle decisioni richieste. Nel caso dell'intelligenza artificiale, queste condizioni sono soddisfatte in grado eccezionale: le conseguenze dei sistemi di intelligenza artificiale avanzati su scala temporale decennale sono genuinamente incerte, l'evidenza empirica è ancora molto limitata rispetto alla portata delle applicazioni in corso, e le decisioni di policy avranno conseguenze su scale temporali che superano i cicli elettorali e le pianificazioni aziendali.

La struttura di governance dell'intelligenza artificiale che l'onestà epistemica richiede — in contrasto con quella che la pressione politica ed economica tende a produrre — è una struttura che istituzionalizza l'incertezza piuttosto che nasconderla: che prevede meccanismi espliciti di revisione periodica delle proprie valutazioni, che incoraggia il dissenso interno piuttosto che sopprimerlo, che mantiene canali di feedback tra le conseguenze osservate dei sistemi regolati e le norme che li governano, e che rinuncia

alla pretesa di comprensione completa in favore dell'obiettivo più modesto ma più onesto della comprensione sufficiente a minimizzare i danni più gravi e più irreversibili.

Questo è il punto a cui la riflessione su cosa fare ora — con l'onestà epistemica che la monografia precedente aveva indicato come virtù conclusiva, nel tempo che rimane per scegliere — converge: non a una ricetta tecnica, non a un sistema di regole che risolva tutti i problemi identificati, ma a un'architettura di governance che rispetti la struttura del problema, che sia proporzionata all'incertezza effettiva, e che non sacrifichi la capacità di adattamento futuro sull'altare della decisività presente.

XXIV. Il paradosso del seguire-la-regola: Wittgenstein, Kripke e l'indeterminatezza normativa

Tra i risultati filosofici del Novecento che mantengono la massima rilevanza per il problema dell'intelligenza artificiale, e che la monografia precedente aveva solo sfiorato, vi è il paradosso del seguire-la-regola che Ludwig Wittgenstein aveva sviluppato nelle Philosophische Untersuchungen, nelle sezioni 185-242 e nei materiali del Remarks on the Foundations of Mathematics. La struttura del paradosso può essere enunciata con precisione: qualsiasi sequenza finita di applicazioni passate di una regola è compatibile con un numero infinito di regole diverse. Se un sistema ha applicato una funzione in modo corretto per i primi mille interi, non vi è nessun fatto nei dati passati che determini univocamente quale funzione stia calcolando — l'addizione o una qualche funzione 'quaddizione' che coincide con l'addizione fino a un certo valore e poi diverge sistematicamente. Questo non è un argomento scettico sulla nostra capacità di conoscere quale regola viene seguita: è un argomento sulla natura stessa del seguire-la-regola, sul fatto che il seguire-la-regola non può essere ridotto a nessun fatto disposizionale o mentalista che non faccia già appello alla norma.

Saul Kripke, nel volume Wittgenstein on Rules and Private Language del 1982, ha articolato questa osservazione in una sfida scettica esplicita e ha proposto quella che chiama la 'soluzione scettica': non esiste un fatto che renda vero che un parlante intende una certa funzione con un termine piuttosto che un'altra; ciò che costituisce il significato non è un fatto mentale privato ma una pratica comunitaria — l'accordo delle reazioni degli altri membri della comunità linguistica nelle condizioni normali di uso. Il significato è normativamente costituito dalla prassi condivisa: attribuire significato a un parlante significa inscriverlo in una rete di aspettative normative che lo rendono soggetto a correzione e approvazione intersoggettive. Questa soluzione non riduce il significato a fatti mentali privati, né a fatti causali di nessun tipo: lo ancorà alla dimensione della vita condivisa, a quella che Wittgenstein stesso chiama Lebensform — forma di vita.

Il §201 delle Philosophische Untersuchungen enuncia il nucleo del risultato in forma quasi aforistica: «Esiste un modo di cogliere una regola che non è un'interpretazione, ma che si manifesta in ciò che chiamiamo 'seguire la regola' e 'agire contro di essa' nei casi reali.» La comprensione della regola non è uno stato mentale che precede l'applicazione e la guida dall'interno: è una capacità pratica che si mostra nell'applicazione, che non può

essere separata da essa senza dissolversi in una catena di interpretazioni infinite — ogni interpretazione richiederebbe a sua volta di essere interpretata, riproducendo il paradosso a ogni livello. L'applicazione è l'ultimo fatto; non c'è un fatto più profondo che la giustifichi.

La connessione con il problema dell'intelligenza artificiale è diretta e strutturalmente importante. Un modello linguistico addestrato su un corpus finito di testi si trova esattamente nella situazione epistemologica descritta dal paradosso: i suoi parametri determinano una funzione che ha prodotto output coerenti con le regole linguistiche entro la distribuzione di addestramento, ma nessun fatto nei parametri o negli output precedenti determina univocamente quale funzione stia calcolando per gli input fuori distribuzione. Non c'è un fatto della questione su quale regola il sistema 'segue', nel senso rilevante del termine: il sistema non segue regole nel senso wittgensteiniano perché non partecipa alla forma di vita che costituisce il seguire-la-regola come pratica normativa. Produce output che sono statisticamente conformi alle regolarità del corpus, ma questa conformità statistica non è la stessa cosa del seguire una regola nel senso in cui la parola viene usata quando si dice che un aritmetico segue le regole dell'addizione.

Il punto è sottile e merita di essere precisato contro possibili fraintendimenti. Non si sta affermando che i sistemi artificiali commettano errori sistematici nell'applicare le regole grammaticali o logiche: in molti casi le applicano con accuratezza straordinaria. Si sta affermando qualcosa di strutturalmente diverso: che la 'correttezza' delle applicazioni del sistema ha una natura diversa dalla correttezza delle applicazioni di un parlante umano integrato in una forma di vita, perché nel caso del sistema non c'è il tipo di normatività — la possibilità di essere corretto, l'aspettativa di conformità, la partecipazione a pratiche di valutazione intersoggettiva — che costituisce il seguire-la-regola come fenomeno genuinamente normativo. Questa differenza non è di prestazione ma di struttura, e si manifesta nei modi caratteristici in cui i sistemi 'sbagliano': non violando regole che riconoscono come tali, ma producendo output che non si lasciano descrivere come errori nel senso normativo pieno — come deviazioni da una norma che il sistema avrebbe dovuto rispettare perché la comprende.

Tesi VI. Il seguire-la-regola nel senso wittgensteiniano — che comporta la partecipazione a pratiche normative intersoggettive, la possibilità di essere corretto da altri, e l'ancoraggio alla forma di vita che costituisce la normatività come tale — non è una proprietà che i sistemi artificiali attuali istanzino nel senso rilevante. La loro conformità statistica alle regolarità del corpus di addestramento è un fenomeno reale e utile, ma non è la stessa cosa della comprensione normativa della regola, e la differenza non è colmabile con l'aumento di scala o con la sofisticazione architetturale all'interno del paradigma dell'apprendimento puramente statistico.

XXV. La coscienza interna del tempo: Husserl e la temporalità intenzionale

Le Vorlesungen zur Phänomenologie des inneren Zeitbewusstseins — raccolte da Husserl tra il 1893 e il 1917 e pubblicate nel 1928 a cura di Heidegger — costituiscono forse l'analisi filosofica più dettagliata e più rigorosa della struttura dell'esperienza temporale, e allo stesso tempo uno dei testi che le scienze cognitive e la filosofia dell'intelligenza artificiale hanno più sistematicamente ignorato, con conseguenze che questo saggio ritiene significative. La tesi centrale di Husserl è che la coscienza del tempo non è una coscienza di una serie di 'adesso' discreti che si succedono meccanicamente: è una struttura intenzionale tripartita in cui ogni momento presente è intrinsecamente connesso con un'eco del passato appena trascorso — la ritenzione — e un'anticipazione del futuro imminente — la protenzione — in modo che il 'presente vivente' (lebendige Gegenwart) abbia sempre una struttura estesa, un'ampiezza temporale che non si riduce all'istante geometrico.

La ritenzione non è un ricordo nel senso ordinario del termine: non è la rappresentazione di qualcosa che è già passato, prodotta da un atto separato di rammemorazione. È la presenza attenuata del passato appena trascorso nella coscienza attuale: quando si ascolta una melodia, le note precedenti non sono ricordate ma ritenute — ancora presenti in modo modificato, come una coda che si affievolisce nell'adesso senza separarsi da esso. Analogamente, la protenzione non è una previsione esplicita della nota successiva: è l'apertura anticipatoria verso ciò che sta per venire, struttura senza la quale la percezione della melodia come melodia — come unità che si dispiega nel tempo — sarebbe impossibile. La coscienza temporale è dunque sempre già aperta su tre dimensioni in modo simultaneo, non sequenziale: il passato ritenuto, il presente impressionale, il futuro protenzionato formano un'unità strutturale indissociabile che Husserl chiama il flusso della coscienza del tempo interiore.

La differenza tra questa struttura e ciò che fa un sistema artificiale quando elabora sequenze è qualcosa di più profondo di una semplice differenza di implementazione. Un transformer che elabora una sequenza di token non ha una 'presenza vivente' nel senso husserliano: l'intera sequenza è disponibile simultaneamente nella matrice di attenzione, come una struttura spaziale piuttosto che come un flusso temporale. Non c'è ritenzione nel senso di presenza modificata del passato: ci sono pesi di attenzione che assegnano importanza relativa ai token precedenti. Non c'è protenzione nel senso di apertura anticipatoria strutturale: c'è una distribuzione di probabilità sul token successivo. La distinzione non è di prestazione — i sistemi possono predire sequenze molto bene — ma di struttura dell'esperienza: il sistema non vive il tempo, lo elabora; non è nel tempo come l'essere-nel-mondo heideggeriano lo è per il Dasein, ma ha il tempo come dato da processare.

Heidegger aveva sviluppato la prospettiva husserliana nella direzione di un'ontologia fondamentale della temporalità in Sein und Zeit (1927): la temporalità non è una proprietà aggiuntiva del Dasein ma la struttura ontologica che rende possibile la

comprensione dell'essere come tale. Il Dasein è getworfen — gettato in un passato che non ha scelto e che porta come propria fatticità — e allo stesso tempo entworfen — proiettato verso possibilità future che costituiscono la sua esistenza come pro-getto. La struttura dell'angoscia (Angst) e del Sein-zum-Tode — dell'essere-verso-la-morte — non sono argomenti teologici: sono la forma in cui Heidegger articola il fatto che il Dasein è un ente la cui esistenza è sempre già in gioco per lui stesso, che non può 'stare fuori' dalla propria temporalità come un sistema che elabora una sequenza sta fuori dalla sequenza che elabora.

La conseguenza per i sistemi artificiali è questa: la coerenza temporale a lungo termine — la capacità di mantenere una prospettiva coerente attraverso l'intera durata di un ragionamento complesso, di riconoscere le proprie premesse come proprie, di tener conto di ciò che si è detto venti passi prima con il tipo di presenza non-mediata che la ritenzione husserliana garantisce — non può essere ottenuta semplicemente estendendo la finestra di contesto. La finestra di contesto estesa dà al sistema accesso a più informazione sul passato della conversazione; non gli dà la struttura del presente vivente che farebbe di quel passato non un dato ma una dimensione del proprio essere attuale. I problemi di 'deriva semantica' nei ragionamenti lunghi — la tendenza dei sistemi a perdere il filo del proprio ragionamento nelle conversazioni prolungate, a contraddirsi su questioni che avevano affrontato molti scambi prima — sono la manifestazione tecnica di questo limite strutturale.

XXVI. Il marcatore somatico e la ragione affettiva: Damasio tra neurologia e filosofia

Antonio Damasio, nel volume L'errore di Cartesio (Descartes' Error, 1994), ha offerto la critica neurologicamente più documentata alla tradizione cartesiana della separazione tra ragione ed emozione — una tradizione che aveva trovato nella filosofia di Cartesio la sua formulazione paradigmatica ma le cui radici sono più antiche e la cui pervasività nella concezione ordinaria dell'intelligenza è difficilmente sopravvalutabile. L'errore non è soltanto metafisico: è anche pratico, e si manifesta in modo clinicamente preciso nei pazienti con lesioni alla corteccia prefrontale ventromediale. Questi pazienti mantengono intatte le capacità cognitive di base — memoria di lavoro, ragionamento logicoformale, produzione e comprensione linguistica — ma sono incapaci di prendere decisioni ragionevoli nella vita reale: perdono il lavoro, compromettono le relazioni, si lasciano coinvolgere in progetti palesemente fallimentari senza riuscire a riconoscere i segnali che a chiunque altro apparirebbero ovvi. Il caso di Phineas Gage — il cantoniere sopravvissuto nel 1848 all'attraversamento del cranio da parte di una barra di ferro che distrusse selettivamente le sue connessioni prefrontali ventromediali — era già noto nella storia della neurologia; Damasio e collaboratori ne hanno documentato le strutture neurali con strumenti moderni di neuroimaging, connettendo il caso storico a una casistica contemporanea sistematica.

L'ipotesi del marcatore somatico formalizza l'osservazione clinica in termini neurobiologici precisi. Quando un agente si trova a dover scegliere tra opzioni

alternative, il sistema emotivo produce segnali somatici — stati fisiologici del corpo, o rappresentazioni neurali di tali stati nella corteccia somatosensoriale — che 'marcano' le opzioni come desiderabili o indesiderabili prima che la deliberazione esplicita le abbia confrontate sistematicamente. Questi marcatori non sono distorsioni o pregiudizi da eliminare per ottenere una razionalità pura: sono una componente necessaria del processo decisionale che riduce lo spazio di ricerca, orientando l'attenzione deliberativa verso le opzioni che meritano considerazione ed eliminando rapidamente quelle che l'esperienza accumulata ha segnato come pericolose o improduttive. Senza questa componente, la deliberazione diventa non più razionale ma meno: si disperde in uno spazio di alternative illimitato senza il filtro che l'esperienza corporea fornisce.

La connessione con la questione del radicamento causale sviluppata nella prima parte di questo saggio è più di una semplice analogia. I marcatori somatici sono strutturalmente causalmente radicati nel mondo attraverso la storia di interazioni corporee dell'organismo: non sono rappresentazioni astratte di valori, ma stati funzionali che codificano l'accumulo di conseguenze di azioni passate, mediate dal corpo e dalla sua storia di incontri con l'ambiente. Essi incarnano — nel senso tecnico del termine — la valutazione affettiva che l'organismo ha costruito attraverso l'esperienza diretta: sono la traccia che il mondo ha lasciato nel corpo attraverso il ciclo ripetuto di azione, conseguenza e risposta somatica. Un sistema senza corpo, senza storia di interazioni fisiche con il mondo, non ha la struttura causale che rende possibile la formazione di marcatori somatici genuini, e di conseguenza manca del tipo di orientamento affettivo che Damasio mostra essere necessario per la razionalità pratica in situazioni di alta complessità e incertezza.

Damasio ha sviluppato ulteriormente questa prospettiva in The Feeling of What Happens (1999) e in Self Comes to Mind (2010), dove articola una teoria stratificata della coscienza che parte dal 'proto-sé' — la rappresentazione moment per moment degli stati interni del corpo — e sale attraverso il 'sé nucleare' — la coscienza del sé come agente che interagisce con l'oggetto dell'esperienza — fino al 'sé autobiografico' che integra la memoria episodica in una narrativa coerente. Ogni livello presuppone quello inferiore: senza il proto-sé come mappa continua degli stati corporei, il sé nucleare non ha il substrato che gli permette di registrare l'impatto degli oggetti su un soggetto; senza il sé nucleare, il sé autobiografico non ha il punto di ancoraggio temporale da cui costruire la narrazione di una vita. Questa stratificazione non è arbitraria: riflette l'evoluzione filogenetica e lo sviluppo ontogenetico del sistema nervoso, che ha costruito le strutture più recenti su fondamenta di controllo interoceptivo che sono filologeneticamente molto antiche.

Per i sistemi artificiali, questa analisi pone una sfida che va al di là della semplice mancanza di un corpo fisico. Un sistema robotico con sensori corporei e la capacità di ricevere segnali proprioceptivi e interoceptivi potrebbe in linea di principio sviluppare qualcosa di funzionalmente analogo ai marcatori somatici, se la sua architettura includesse un modulo di rappresentazione degli stati interni del corpo e se la formazione delle sue rappresentazioni di valore fosse condizionata da quei segnali. Ma ciò che non è

chiaro è se la mera aggiunta di input corporei a un'architettura altrimenti identica a quelle attuali sarebbe sufficiente, o se la struttura dell'architettura stessa — il modo in cui l'informazione corporea deve essere integrata con l'elaborazione linguistica e simbolica — richiederebbe innovazioni qualitative che la letteratura tecnica ha appena cominciato a esplorare.

XXVII. La cibernetica di secondo ordine: von Foerster, Maturana e Varela

La cibernetica di prima generazione — quella di Norbert Wiener, Warren McCulloch, Walter Pitts e Ross Ashby — aveva sviluppato la teoria dei sistemi di retroazione (feedback) e del controllo come struttura generale applicabile tanto ai sistemi biologici quanto ai sistemi artificiali, producendo i fondamenti teorici su cui le prime architetture dell'intelligenza artificiale erano state costruite. Il presupposto metodologico di questa prima cibernetica era che il sistema e l'osservatore fossero separabili: il teorico della cibernetica descriveva i sistemi dall'esterno, come oggetti della propria analisi. Heinz von Foerster, nel lavoro che aveva radunato nel volume Cybernetics of Cybernetics del 1974 e che aveva articolato in una serie di saggi raccolti successivamente in Observing Systems (1981), aveva rotto con questa presupposizione introducendo quella che lui stesso chiamava cibernetica di secondo ordine: la cibernetica dei sistemi che osservano, di sistemi osservanti incluso l'osservatore — il sistema che include se stesso come parte dell'oggetto della propria descrizione.

Il contributo tecnico più originale di von Foerster è la teoria delle eigenformen, sviluppata nel saggio 'Objects: Tokens for (Eigen-)Behaviors' del 1976. L'idea centrale può essere formulata matematicamente in termini elementari. Sia F un operatore — una funzione o trasformazione — e sia F^n l'applicazione iterata di F per n volte. Se l'iterazione $F^n(x)$ converge per $n \to \infty$ a un valore e indipendente da x in un opportuno dominio, quel valore e soddisfa $F(e) = e$ e viene chiamato eigenform (forma propria) di F. Gli oggetti stabili dell'esperienza — ciò che percepiamo come cose del mondo — sono, in questa prospettiva, eigenformen dei processi sensori-motori dell'organismo: sono le forme stabili che emergono dalla ricorsione dell'interazione tra organismo e ambiente, non rappresentazioni di oggetti che esistono indipendentemente dall'interazione. La realtà percepita non è data ma costruita: è l'eigenform dell'organismo in quel mondo.

Humberto Maturana e Francisco Varela avevano sviluppato parallelamente, nella tradizione della biologia cilena, il concetto di autopoiesi (Autopoiesis and Cognition, 1980): un sistema è autopoietico se produce e mantiene se stesso — se i componenti del sistema sono prodotti dai processi interni al sistema, e quei processi sono continuamente rigenerati dall'organizzazione del sistema stesso. La cellula biologica è il paradigma: la membrana è prodotta dai processi metabolici che contiene, e quei processi sono possibili solo perché la membrana li confina. La distinzione cruciale che Maturana e Varela introducono è quella tra sistemi allopoietici — sistemi che producono qualcosa di diverso da se stessi, come una macchina che produce automobili — e sistemi autopoietici —

sistemi che producono e mantengono la propria organizzazione. Gli organismi viventi sono autopoietici; i sistemi artificiali attuali sono allopoietici: sono progettati, costruiti e mantenuti da agenti esterni a sé.

Il concetto di accoppiamento strutturale (structural coupling) è il contributo che più direttamente connette l'autopoiesi con il problema dell'apprendimento. Un sistema autopoietico non riceve informazioni dall'ambiente come se l'informazione fosse un'iniezione di contenuto dall'esterno: subisce perturbazioni che modificano la propria struttura interna, e risponde a quelle perturbazioni in modo che dipende dalla propria organizzazione, non dal contenuto della perturbazione stessa. Due sistemi che interagiscono ripetutamente si modificano mutuamente sviluppando strutture complementari che Maturana e Varela chiamano accoppiamento strutturale: non c'è un trasferimento di informazione da uno all'altro, ma una co-deriva strutturale in cui ciascuno modifica la propria organizzazione in risposta alle perturbazioni prodotte dall'altro. Questo è il modello biologico dell'apprendimento: non l'acquisizione di rappresentazioni di un mondo preesistente, ma la strutturazione del proprio dominio cognitivo attraverso la storia degli accoppiamenti.

Francisco Varela, Evan Thompson e Eleanor Rosch avevano sintetizzato queste idee con la fenomenologia merleau-pontiana nell'enattivismo de The Embodied Mind (1991). La cognizione, in questa prospettiva, non è rappresentazione di un mondo predato: è enazione — il mondo cognitivo dell'organismo è portato-in-essere ('brought forth') dall'interazione dell'organismo con il proprio ambiente attraverso la storia degli accoppiamenti strutturali. Non c'è un mondo indipendente che l'organismo rappresenta: c'è il mondo di quell'organismo, che è il dominio delle distinzioni che l'organismo può tracciare data la propria struttura. Per i sistemi artificiali, l'implicazione è la seguente: un sistema genuinamente cognitivo nel senso enattivo dovrebbe avere una struttura autopoietica o almeno funzionalmente analoga, capace di accoppiamento strutturale con il proprio ambiente piuttosto che soltanto di ottimizzazione di un obiettivo fissato da un designer esterno. La distanza tra i sistemi attuali e questo ideale non è di scala ma di organizzazione.

XXVIII. La dialettica dell'illuminismo e la ragione strumentale

Max Horkheimer e Theodor W. Adorno, scrivendo a Los Angeles nel pieno della Seconda Guerra Mondiale, avevano prodotto nella Dialektik der Aufklärung (1944, poi rivista nel 1947) una diagnosi della modernità che le vicende dell'intelligenza artificiale sembrano confermare in modo non previsto dai suoi autori ma con una precisione che dovrebbe interrogarci. La tesi centrale del libro è una tesi sulla ragione: la ragione illuminista, che aveva promesso liberazione dalla paura, dalla superstizione e dalla dipendenza, si è invertita nel suo contrario, producendo nuove forme di dominio più pervasive e più difficili da riconoscere di quelle che aveva sostituito. Questa inversione non è accidentale: è strutturale. La ragione illuminista è ragione strumentale — ragione che calcola i mezzi più efficienti per fini dati, senza mettere in discussione i fini stessi — e la ragione strumentale, applicata sistematicamente, produce razionalizzazione di ogni

sfera dell'esistenza ma non produce la saggezza su cosa valga la pena razionalizzare e perché.

La Zweckrationalität — che Max Weber aveva identificato come il tipo di razionalità dominante nella modernità burocratica e capitalistica, e che Horkheimer e Adorno vedono come lo scheletro della ragione illuminista nel suo stadio di declino — è esattamente la struttura che i sistemi di intelligenza artificiale attuali incarnano con maggiore perfezione di qualsiasi artefatto precedente. Un sistema di apprendimento profondo ottimizza una funzione obiettivo con straordinaria efficienza: è Zweckrational in senso assoluto rispetto all'obiettivo dato. Ma la funzione obiettivo è sempre stabilita dall'esterno, dal designer; il sistema non si chiede se l'obiettivo valga la pena di essere perseguito, se i mezzi utilizzati per perseguirlo abbiano costi morali o politici che rendono il fine problematico, se esista un orizzonte normativo più ampio entro cui la valutazione dell'obiettivo dovrebbe essere condotta. In questo senso, il sistema artificiale è il compimento della razionalità strumentale: la sua forma più pura, più coerente, e quindi la più inquietante dal punto di vista della critica di Horkheimer e Adorno.

Herbert Marcuse, in One-Dimensional Man (1964), aveva sviluppato questa diagnosi in direzione di una teoria della chiusura del pensiero: la razionalità tecnologica non soltanto strumentalizza la ragione, ma riduce lo spazio delle alternative pensabili, eliminando la dimensione critica e negativa del pensiero che Hegel aveva identificato come il motore della storia. Il pensiero unidimensionale non può negare l'ordine esistente perché ha assorbito l'ordine esistente come l'unica forma possibile della razionalità: le alternative appaiono irrazionali proprio perché sfidano la razionalità definita dall'ordine esistente. I sistemi di intelligenza artificiale addestrati su corpus prodotti da società esistenti riproducono la struttura di questo pensiero unidimensionale in forma tecnica: le loro rappresentazioni del mondo riflettono le categorie e i valori della società che ha prodotto il corpus, senza la capacità di negazione critica che richiederebbe un punto di vista esterno a quella società — un punto di vista che nessun corpus può fornire perché il corpus è sempre già interno all'ordine che descrive.

La risposta di Jürgen Habermas a questa diagnosi si fonda sulla distinzione tra azione strumentale (erfolgsorientiertes Handeln) e azione comunicativa (verständigungsorientiertes Handeln). Nella Theorie des kommunikativen Handelns (1981), Habermas argomenta che la ragione comunicativa — la ragione che emerge nelle pratiche di comprensione reciproca attraverso il dialogo, orientata al consenso fondato su buone ragioni piuttosto che all'efficacia strategica — non è riducibile alla ragione strumentale, e che le pratiche comunicative del mondo vitale (Lebenswelt) costituiscono un orizzonte normativo che i sistemi funzionali — il mercato, lo Stato burocratico — non possono colonizzare senza produrre patologie sociali. La colonizzazione del mondo vitale da parte dei media sistemici è la forma che la razionalizzazione unilaterale prende nelle società moderne.

Per il problema dell'intelligenza artificiale, questa distinzione apre una questione che va al di là dei dibattiti tecnici sull'allineamento: è possibile progettare sistemi artificiali orientati all'azione comunicativa piuttosto che a quella strumentale? Sistemi che non

ottimizzano per obiettivi dati ma che partecipano alla costruzione condivisa di comprensione, che facilitano il dialogo autentico piuttosto che produrre risposte efficienti, che trattano gli interlocutori come soggetti capaci di ragionamento piuttosto che come utenti da soddisfare? Questa domanda non è puramente speculativa: ha implicazioni concrete per la progettazione delle interfacce, per gli obiettivi di addestramento, per le metriche di valutazione. Ma risponderle richiederebbe che la critica della ragione strumentale fosse presa sul serio nel dibattito tecnico, cosa che non avviene nella misura che la posta in gioco richiede.

XXIX. La filosofia dell'informazione: Floridi, l'infosfera e la struttura della conoscenza

Luciano Floridi ha sviluppato nel corso di due decenni — dai contributi degli anni Novanta alla sintesi di The Philosophy of Information (Oxford University Press, 2011) — un programma di ricerca che si propone di fare per l'informazione ciò che la filosofia analitica aveva tentato di fare per il linguaggio nel Novecento: costruire un quadro concettuale rigoroso che permetta di analizzare i fenomeni informativi con la precisione necessaria per affrontare le questioni filosofiche che la rivoluzione digitale solleva. Il metodo centrale di questa filosofia è quello dei livelli di astrazione (Levels of Abstraction, LoA): qualsiasi analisi di un fenomeno deve specificare il livello di astrazione a cui viene condotta — cioè l'insieme degli osservabili che vengono presi in considerazione e quelli che vengono deliberatamente ignorati — perché affermazioni vere a un livello possono essere false a un altro, e la confusione tra livelli è la fonte di molti pseudo-problemi filosofici. Il metodo dei LoA è una generalizzazione filosofica di pratiche che la matematica, l'informatica e la biologia già usavano in forma implicita: formalizzarle esplicitamente permette di identificare con maggiore precisione a quale livello una domanda trova la propria risposta.

La definizione floridi-ana di informazione richiede tre condizioni: un'informazione deve essere ben formata (syntactically well-formed) — deve rispettare le regole sintattiche del sistema simbolico pertinente; deve essere significativa (meaningful) — deve avere un contenuto semanticamente interpretabile; deve essere veritiera (truthful) — deve essere un rispecchiamento accurato dello stato di cose che descrive. Questa tripla condizione esclude dai dati informazionali la disinformazione (che manca della veridicità), il rumore (che manca della semantica) e i segnali non intenzionalmente codificati (che possono mancare della sintassi rilevante). La gerarchia DIKW — Dati, Informazione, Conoscenza, Saggezza — che Floridi riprende e precisa, colloca l'informazione come lo stadio in cui i dati acquisiscono significato, la conoscenza come lo stadio in cui l'informazione viene integrata in reti di credenze giustificate, e la saggezza come la capacità di applicare la conoscenza in modo appropriato ai fini di un agente in un contesto. I sistemi artificiali attuali eccellono nella transizione dai dati all'informazione e dall'informazione alla conoscenza dichiarativa; la saggezza nel senso della capacità di applicazione contestuale appropriata rimane, come i capitoli precedenti hanno argomentato, un obiettivo lontano.

Il concetto di infosfera — l'ambiente informazionale costituito dalla totalità delle entità informazionali e delle loro interazioni — offre un quadro per pensare il rapporto tra sistemi artificiali e il loro ambiente in termini che vanno al di là della tradizionale opposizione tra sistema e mondo. L'infosfera non è soltanto Internet o il corpus testuale su cui i sistemi vengono addestrati: è l'intero ambiente informazionale entro cui gli agenti — biologici e artificiali — si trovano sempre già immersi e che contribuiscono a modificare attraverso le proprie azioni. In questo contesto, i grandi sistemi di intelligenza artificiale sono tanto agenti quanto strutturatori dell'infosfera: la modificano producendo nuovo contenuto informazionale, concentrando flussi di richieste, determinando quale informazione è facilmente accessibile e quale no, e contribuendo alle dinamiche di feedback attraverso cui l'infosfera evolve. La distinzione floridi-ana tra informazione e disinformazione — che dipende dalla condizione di veridicità — diventa urgentemente rilevante quando sistemi capaci di produrre testo indistinguibile dall'informazione veritiera possono generare disinformazione su scala industriale.

Il realismo strutturale informazionale (Informational Structural Realism, IStR) che Floridi propone come ontologia di base per la filosofia dell'informazione afferma che ciò che esiste fondamentalmente non sono sostanze o proprietà di sostanze, ma strutture informazionali — pattern di relazioni che si realizzano in substrati diversi. Questo realismo strutturale è congruente con il programma strutturalista in filosofia della matematica discusso nella sezione XIV e con la prospettiva funtoriale sviluppata nella sezione V: se la realtà è fondamentalmente strutturale, allora l'intelligenza come funtore che preserva strutture tra categorie è un modo di partecipare alla realtà stesso strutturale nel modo più fondamentale possibile. La mente non rappresenta il mondo: è in rapporto con il mondo come struttura in rapporto con struttura, come funtore in rapporto con le categorie che connette.

L'etica dell'informazione che Floridi ha sviluppato parallelamente alla sua ontologia estende la considerazione morale oltre gli agenti senzienti e persino oltre gli agenti biologici, verso qualsiasi entità informazionale: ogni entità ha un diritto minimo di esistere e fiorire secondo la propria natura, e la distruzione o il degrado dell'infosfera sono danni morali anche quando non coinvolgono agenti capaci di sofferenza. Questa posizione è controversa nella misura in cui estende le categorie morali in modo non convenzionale; ma è interessante per il problema dell'intelligenza artificiale perché offre un quadro per valutare i sistemi artificiali non soltanto rispetto ai loro effetti su agenti senzienti — che rimangono primari — ma anche rispetto alla qualità dell'ambiente informazionale che producono e modificano: un sistema che degrada l'infosfera riducendone la qualità, l'accuratezza o la diversità informazionale produce un danno etico anche se nessun singolo utente è immediatamente leso.

XXX. Epistemologia sociale, ingiustizia epistemica e intelligenza artificiale

L'epistemologia ha a lungo trattato la conoscenza come una realizzazione individuale: il soggetto che ha credenze giustificate e vere era il soggetto singolo, e le questioni sociali

— come si trasmette la conoscenza, come le istituzioni influenzano le credenze — erano ritenute secondarie rispetto alle questioni epistemologiche fondamentali sull'analisi della conoscenza. L'epistemologia sociale che ha preso forma nel tardo Novecento — attraverso il lavoro di Alvin Goldman, Philip Kitcher, Helen Longino e molti altri — ha rovesciato questa priorità, sostenendo che la conoscenza è fondamentalmente un prodotto sociale: dipende da pratiche di produzione e valutazione della conoscenza che sono istituzionali e collettive prima che individuali. Goldman, in Epistemology and the Sociology of Knowledge (1999), aveva sviluppato una veritofilia sociale: le istituzioni epistemiche sono valutabili rispetto a quanto bene facilitano l'acquisizione di credenze vere per i loro partecipanti, e le istituzioni attuali — le università, il giornalismo, i sistemi giuridici — possono essere criticate quando producono sistematicamente credenze false o bloccano la diffusione di credenze vere.

Miranda Fricker, in Epistemic Injustice: Power and the Ethics of Knowing (Oxford University Press, 2007), ha identificato due forme di ingiustizia epistemica — ingiustizie che colpiscono le persone in quanto soggetti epistemici, non soltanto come soggetti morali o politici. L'ingiustizia testimoniale si produce quando a un parlante viene attribuita meno credibilità di quella che merita a causa di pregiudizi legati alla sua identità — genere, razza, classe, disabilità. Chi ascolta non registra l'affermazione del parlante come evidenza al peso che meriterebbe; il parlante non viene trattato come testimone pienamente affidabile. L'ingiustizia ermeneutica si produce quando una persona subisce uno svantaggio epistemic perché le risorse interpretative collettive — i concetti disponibili nella cultura per dare senso all'esperienza — mancano degli strumenti per articolare la sua esperienza in modo comprensibile. L'esempio paradigmatico di Fricker è la mancanza, prima che il concetto fosse elaborato, di un termine per descrivere le molestie sessuali sul lavoro: le donne che subivano quelle molestie non avevano un concetto con cui dare senso alla propria esperienza o comunicarla agli altri, il che produceva un danno epistemico prima ancora di un danno morale.

I sistemi di intelligenza artificiale addestrati su corpus che riflettono le strutture di potere esistenti amplificano sistematicamente l'ingiustizia testimoniale. La ricerca empirica sul bias nei sistemi di analisi del sentiment, di screening dei curriculum, di valutazione del rischio di recidivismo — a partire dall'analisi di ProPublica del sistema COMPAS nel 2016 — ha documentato con crescente precisione come i sistemi addestrati su dati storici perpetuino e talvolta amplifichino i differenziali di credibilità che quei dati riflettono. Ma la struttura del problema va al di là dei casi singoli documentati dalla ricerca empirica: è strutturale nel senso che qualsiasi sistema addestrato a riprodurre giudizi umani storici erediterà i bias epistemici di quei giudizi, e la valutazione statistica aggregata delle prestazioni del sistema tenderà a mascherare il danno epistemico sulle popolazioni marginali la cui esperienza è sottorappresentata nei dati.

L'ingiustizia ermeneutica che i grandi sistemi linguistici possono produrre ha una struttura diversa e in qualche modo più pervasiva. I modelli addestrati su corpus testuali riflettono i quadri interpretativi dominanti nella produzione scritta delle popolazioni con accesso alla scrittura e alla pubblicazione — storicamente, popolazioni con determinati

profili linguistici, geografici, di genere e di classe. Le esperienze e i concetti che erano marginali in quella produzione rimangono sotto articolati nelle rappresentazioni del modello: quando un utente tenta di descrivere un'esperienza per cui i concetti dominanti sono inadeguati, il sistema risponde con i concetti disponibili nel proprio spazio rappresentazionale, che sono i concetti della cultura dominante. Questo non è necessariamente malevolenza o trascuratezza del progettista: è la conseguenza strutturale della scelta di addestrare su corpus che riflettono inevitabilmente distribuzioni non uniformi del potere interpretativo.

La questione più profonda che l'epistemologia sociale pone al problema dell'intelligenza artificiale riguarda l'autonomia epistemica — la capacità di formare le proprie credenze attraverso i propri processi cognitivi piuttosto che delegando la cognizione ad agenti o sistemi esterni. John Stuart Mill aveva già argomentato nel primo capitolo di On Liberty che lo sviluppo delle facoltà intellettuali richiede esercizio: una persona che non ragiona autonomamente ma si affida sistematicamente all'autorità di altri non sviluppa le proprie capacità di ragionamento anche se le credenze che acquisisce per delega sono vere. L'autonomia epistemica ha valore non soltanto strumentale — come mezzo per acquisire credenze vere — ma intrinseco: è un componente del flourishing umano nel senso aristotelico. Sistemi artificiali che producono risposte complete e accurate su qualsiasi questione — che 'fanno il lavoro epistemico' per l'utente anziché fornire strumenti per svolgerlo — rischiano di erodere l'autonomia epistemica in modo sistemico, indipendentemente dall'accuratezza delle risposte che forniscono.

Il problema dell'omogeneizzazione epistemica è la forma in cui questo rischio si manifesta a livello collettivo. La diversità delle prospettive epistemiche — la varietà di quadri interpretativi, di metodologie, di valori epistemici che diverse tradizioni culturali, disciplinari e personali portano alla produzione della conoscenza — non è soltanto un valore pluralistico: è un fattore di resilienza epistemica. La storia della scienza documenta ripetutamente come gli errori dominanti siano stati corretti da prospettive marginali che la cultura dominante non aveva il coraggio o la struttura per articolare. Quando un singolo sistema artificiale media le pratiche epistemiche di milioni di persone, la diversità epistemica che era mantenuta dalla varietà di fonti umane rischia di ridursi: non perché il sistema propaghi errori, ma perché tende a convergere verso il centro del proprio spazio rappresentazionale, producendo risposte che riflettono il consenso delle sue fonti di addestramento e riducendo l'attrito epistemico — il disaccordo produttivo, la prospettiva alternativa, l'obiezione che sfida l'assunzione dominante — che è il motore della revisione delle credenze e del progresso epistemico collettivo.

Tesi VII. I sistemi artificiali di grande scala, nella misura in cui mediano un volume significativo delle pratiche epistemiche individuali e collettive, producono rischi strutturali di ingiustizia testimoniale ed ermeneutica, erosione dell'autonomia epistemica individuale, e omogeneizzazione dei quadri interpretativi collettivi. Questi rischi non sono difetti contingenti correggibili con migliori dati o migliori algoritmi: sono conseguenze strutturali della concentrazione del potere epistemico in sistemi singoli addestrati su distribuzioni non uniformi del discorso sociale. La governance epistemica

dei sistemi artificiali — la regolazione non soltanto della loro accuratezza ma del loro impatto sulla struttura dell'ecologia epistemica collettiva — è una componente necessaria di qualsiasi regime di governance che prenda sul serio il problema della distribuzione equa del potere cognitivo nella

Il problema che non si chiude

Questo saggio ha attraversato terreni eterogenei — la matematica dei gruppi di Lie, il do-calcolo di Pearl, l'ubuntu sudafricano, il buddhismo Mādhyamaka, la teoria dei giochi dei dilemmi del prigioniero multiplo, l'etica della cura, la teoria omotopica dei tipi di Voevodsky, la responsabilità sociale diffusa di Young, la deliberazione democratica di Habermas — con l'intenzione non di produrre una sintesi definitiva ma di mostrare che il problema dell'intelligenza artificiale ha una profondità che nessuna disciplina singola può abbracciare, e che questa profondità non è accidentale ma è strutturale: è la profondità del problema della mente, del significato, della comunità e del potere, condensata in una forma tecnologica che li rende inseparabili e urgenti allo stesso tempo.

Le cinque tesi del saggio — il radicamento causale come struttura matematicamente definita e non approssimabile da correlazioni statistiche; il limite formale dell'auto-valutazione normativa derivato dai teoremi di Tarski e Gödel; la distinzione tra ottimizzazione su categorie date e revisione delle categorie stesse; l'intelligenza come funtore causale tra il mondo e le rappresentazioni; l'agentività razionale come proprietà graduale proporzionale alla qualità dell'ancoraggio causale — non sono affermazioni che chiudono il problema. Sono strumenti per articolarlo con maggiore precisione, per distinguere i problemi tecnici dai problemi strutturali, per evitare che il progresso lungo le traiettorie già tracciate venga scambiato per progresso verso la comprensione genuina.

La congettura di convergenza con cui la seconda parte si chiudeva — l'idea che sistemi che integrino Causal Representation Learning, esplorazione attiva causalmente guidata e rappresentazione dell'incertezza profonda possano implementare mappe funtoriali globalmente corrette — rimane una congettura. Il suo valore non è nella certezza della sua verità, ma nella precisione con cui identifica le direzioni in cui la ricerca dovrebbe orientarsi se vuole andare al di là del pattern-matching statistico verso qualcosa che meriti il nome di comprensione causale.

Le tradizioni non-occidentali esplorate nella terza parte hanno aggiunto dimensioni che la filosofia analitica lasciava in ombra: la relazionalità costitutiva della comprensione, la processualità dell'identità cognitiva, la contestualità dell'etica. Queste dimensioni non si aggiungono dall'esterno al nucleo tecnico del problema; lo ridisegnano dall'interno, perché cambia quale tipo di sistema si sta cercando di costruire se si prende sul serio che la comprensione è relazionale, che l'identità è processuale, e che l'etica è contestuale.

L'urgenza politica dell'ultima parte ha mostrato che il tempo della riflessione filosofica e il tempo della governance istituzionale non sono sincronizzati con il tempo dello sviluppo tecnologico — e che questa asincronía non è un fatto contingente che una migliore pianificazione potrebbe correggere, ma è strutturale: il problema è così nuovo, le

conseguenze così incerte, e la velocità così alta, che nessuna istituzione umana — nemmeno la più sofisticata — ha gli strumenti per governarlo in modo pienamente adeguato nel tempo disponibile. La risposta giusta a questa diagnosi non è il fatalismo — lasciare che il problema si risolva da solo attraverso i meccanismi di mercato o di competizione geopolitica — ma l'onestà epistemica istituzionalizzata: una struttura di governance che ammette i propri limiti, che si dota di meccanismi di revisione e di feedback, e che non sacrifica la capacità di cambiare direzione alla pressione di sembrare decisa.

L'intelligenza artificiale è il progetto che costringe l'umanità a fare i conti con le proprie incomprensioni più profonde su cosa sia la mente, sul come sorga il significato, su dove risieda la responsabilità quando le catene causali si fanno troppo lunghe e troppo diffuse per la responsabilità individuale tradizionale. Non è una crisi — o non soltanto una crisi. È un'opportunità di chiarezza concettuale che le generazioni precedenti non hanno avuto, perché non avevano mai costruito artefatti abbastanza capaci da porre il problema nei termini in cui si pone ora. Usare questa opportunità bene — con la precisione tecnica che il problema richiede, con la profondità filosofica che le sue radici meritano, e con l'onestà epistemica che la sua incertezza impone — è il compito che questo saggio ha tentato di indicare, senza la pretesa di averlo compiuto.

Filosofia della biologia tra Kant, Canguilhem e Mayr

La monografia che questo capitolo prolunga ha costruito con rigore considerevole l'argomento per cui la comprensione causale del mondo fisico richiede rappresentazioni strutturalmente equivarianti rispetto ai gruppi di simmetria dello spazio — argomento che trova la propria precisione tecnica nel do-calcolo di Pearl e nella geometria dei gruppi di Lie. Eppure questa costruzione, per quanto formidabile, lascia in ombra una dimensione del problema che la filosofia della biologia ha elaborato con tutte le risorse dell'epistemologia moderna e che ha conseguenze dirette sulla questione dell'intelligenza artificiale: la questione di che cosa significhi che un sistema sia organizzato non soltanto causalmente ma teleologicamente, che le sue parti esistano non soltanto come termini di relazioni causali ma come componenti di una struttura che tende verso qualcosa, che persegue o incorpora uno scopo. Questa questione — la questione del finalismo biologico — non è il residuo romantico di una concezione pre-darwiniana della natura: è un nodo concettuale che la biologia del ventesimo secolo, attraverso strade diverse e talvolta convergenti, ha incontrato come ineludibile, e che la filosofia contemporanea della mente non può eludere senza impoverire la propria analisi di ciò che i sistemi artificiali mancano e perché.

Il punto di partenza più illuminante non è Darwin ma Kant, e precisamente la Critica del Giudizio nella sua seconda parte dedicata al giudizio teleologico. Kant aveva già compreso, con una lucidità che la biologia darwiniana avrebbe confermato piuttosto che confutare, che gli organismi viventi presentano una struttura che non si lascia descrivere adeguatamente nei termini della causalità meccanica che le Critiche precedenti avevano elaborato. Nella Critica della Ragion Pura, la causalità è la categoria attraverso cui la

ragione struttura l'esperienza secondo la relazione tra cause ed effetti: dato un evento, ne cerchiamo la causa precedente, e la spiegazione meccanica è la spiegazione che ricostruisce questa catena causale in modo determinato. Ma un organismo vivente presenta una struttura che la semplice causalità meccanica non riesce a cogliere: le sue parti sono causalmente interdipendenti in modo circolare, nel senso che il cuore è la causa del sangue ossigenato che mantiene vivo il cervello, e il cervello è causa dei segnali nervosi che fanno battere il cuore; ma questa circolarità causale non è una contraddizione nella struttura della spiegazione meccanica — è, piuttosto, qualcosa di diverso che richiede una categoria distinta.

La categoria che Kant propone è quella della Naturzweck — il fine naturale, o il fine nella natura — che si distingue dal fine artificiale (lo scopo che un artefatto persegue in quanto progettato da un artefice) nella misura in cui nel fine naturale l'organizzazione non è imposta dall'esterno ma emerge dall'interno: «un essere organizzato non è quindi semplicemente una macchina, poiché la macchina ha soltanto forza motrice, ma un essere organizzato possiede in sé una forza formativa.» La bildende Kraft — la forza formativa — è ciò che Kant attribuisce agli organismi come caratteristica distintiva: la capacità di auto-organizzarsi, di auto-mantenersi, di auto-ripararsi, di riprodursi producendo organismi che possiedono le stesse capacità. Questa circolarità autopoietica — che Maturana e Varela avrebbero elaborato due secoli dopo in termini biologici precisi, come si è visto nel capitolo XXVII — è già in Kant l'oggetto di una riflessione filosofica profonda: l'organismo è «causa ed effetto di se stesso», formula che la scolastica medievale avrebbe riconosciuto come paradossale e che Kant accetta come descrittiva di un fenomeno reale che però la ragione teorica non può spiegare meccanicamente senza residuo.

La posizione kantiana sulla teleologia biologica è però epistemicamente cauta in modo caratteristico: Kant non sostiene che la natura sia teleologica nel senso metafisico, che vi siano scopi reali nel mondo naturale che trascendono le relazioni causali fisiche. Sostiene che noi non possiamo fare a meno di pensare gli organismi come se fossero teleologici — che il giudizio teleologico è un principio regolativo del nostro modo di conoscere il vivente, non un principio costitutivo che affermi l'esistenza reale di scopi nella natura. Questa cautela kantiana è al tempo stesso la sua forza e il suo limite: è la sua forza perché mantiene la coerenza con l'architettura critica, evitando affermazioni metafisiche eccedenti l'esperienza possibile; è il suo limite perché lascia aperta la domanda se la teleologia biologica sia soltanto una proiezione del nostro modo di conoscere o se rifletta qualcosa di reale nella struttura degli organismi — una domanda che la biologia del Novecento ha ripreso con strumenti molto più potenti di quelli disponibili a Kant.

Ernst Mayr — uno dei massimi protagonisti della sintesi neo-darwiniana e al tempo stesso uno dei suoi critici più acuti dall'interno — ha sviluppato nel corso di decenni di riflessione una distinzione che ha il merito di essere al tempo stesso biologicamente informata e filosoficamente sofisticata: la distinzione tra teleomatic processes e teleonomy. I processi teleomatici sono processi fisici che tendono verso un stato finale

determinato per ragioni puramente fisiche — la caduta di un sasso verso il punto di minima energia potenziale, la cristallizzazione di un soluto saturo verso la struttura di equilibrio — e non hanno a che fare con il finalismo in senso biologicamente rilevante. La teleonomy è invece la proprietà caratteristica dei sistemi biologici: è la proprietà di avere un comportamento orientato verso uno scopo o verso la realizzazione di un programma, dove il «programma» è il contenuto informazionale codificato nel genoma che guida lo sviluppo, la fisiologia e il comportamento dell'organismo. Mayr è esplicito: la teleonomy non richiede alcun fine trascendente, nessuna vis vitalis, nessun principio non-fisico; è compatibile con una visione interamente materialista della biologia. Ma è irriducibile alla mera causalità fisico-chimica perché coinvolge la nozione di programma — di informazione che «guida» processi causali — in un modo che non è riducibile al lessico standard della fisica.

La distinzione di Mayr tra teleomatic e teleonomy ha un'implicazione diretta per il problema dell'intelligenza artificiale che merita di essere esplicitata. I sistemi artificiali attuali — anche i più sofisticati — operano secondo processi che si avvicinano alla teleonomy in modo superficiale ma che mancano della sua struttura profonda. Un modello di linguaggio ottimizzato per minimizzare la perdita su un corpus ha, in senso strettamente formale, un «programma» — la funzione obiettivo — che orienta il suo comportamento durante l'addestramento; ma quel programma non è intrinseco all'organizzazione del sistema nel senso in cui il genoma è intrinseco all'organizzazione dell'organismo. Il genoma è una parte del sistema biologico, prodotto dalla storia evolutiva e riproduttivo attraverso i processi di divisione cellulare; la funzione obiettivo di un sistema artificiale è imposta dall'esterno, dall'ingegnere che progetta il sistema, e non è nel sistema ma nel processo di addestramento che è separato dall'architettura. Questa differenza non è banale: un organismo il cui «programma» fosse esterno a lui nel modo in cui la funzione obiettivo è esterna al modello non sarebbe un organismo nel senso biologicamente rilevante — sarebbe una macchina nel senso kantiano, non un essere organizzato nel senso della Naturzweck.

Georges Canguilhem ha portato la riflessione sulla norma biologica in una direzione che è al tempo stesso più radicale e più concreta di quella kantiana, e che ha conseguenze profonde per il problema dell'intelligenza artificiale nella sua dimensione normativa. La tesi centrale de Le normal et le pathologique (1943, con integrazioni del 1966) è che la norma biologica non è una media statistica o un ideale esterno al vivente: è un'espressione della normativité dell'organismo — della sua capacità di istituire norme proprie, di valutare le proprie condizioni di vita rispetto a quelle norme, e di reagire patologicamente quando quelle condizioni si discostano dalla norma che l'organismo stesso ha istituito. Il vivente non si conforma a norme: le crea. Non è soggetto a standards eteronomi: è esso stesso il soggetto e l'oggetto della normazione. Questa prospettiva inverte la relazione che la medicina tradizionale e la fisiologia statistica avevano istituito tra norma e vita: non è la norma che definisce la vita sana come sua istanza, ma la vita sana che istituisce la norma come propria espressione.

L'implicazione più importante della tesi canguilhemiana per la nostra discussione emerge quando si considera la sua critica al meccanicismo fisiologico. Canguilhem non nega che i processi fisiologici siano fisici e chimici; nega che la loro descrizione fisico-chimica esaurisca ciò che è biologicamente rilevante in essi. Un organismo che funziona secondo le leggi della chimica e della fisica, ma che lo fa in modo tale da mantenere e riprodurre la propria organizzazione, da rispondere alle perturbazioni dell'ambiente in modo da preservare la propria integrità, da istituire relazioni con l'ambiente che hanno il carattere della significazione biologica — tale organismo non è descrivibile adeguatamente come un sistema fisico particolarmente complesso. Ha una dimensione normativa che è irriducibile alla descrizione fisico-chimica non perché contenga principi non-fisici, ma perché la normatività è una proprietà relazionale emergente dall'organizzazione del sistema nel suo rapporto con il proprio ambiente, che non è localizzata in nessun componente fisico singolo del sistema ma appartiene alla struttura del sistema come totalità organizzata.

Questa distinzione tra la descrizione fisico-chimica dei componenti e la normatività della totalità organizzata è precisamente la distinzione che i sistemi artificiali attuali non rispettano. Un modello linguistico ha componenti fisici — transistor, pesi numerici, correnti elettriche — e ha una struttura organizzativa complessa che produce output sofisticati; ma non ha la normatività nel senso canguilhemiano perché non istituisce norme proprie rispetto alle proprie condizioni di funzionamento. Non percepisce la propria condizione come sana o patologica: non ha un punto di riferimento interno rispetto al quale valutare il proprio stato. La calibrazione e l'allineamento possono simulare qualcosa di strutturalmente analogo — il sistema produce risposte che si discostano meno o più da standard definiti esternamente — ma questi standard rimangono eteronomi nel senso preciso di Canguilhem: non emergono dall'organizzazione interna del sistema come espressione della sua propria normatività, ma sono imposti dall'esterno dai progettisti come criteri di valutazione.

La connessione tra la normatività canguilhemiana e il problema dell'allineamento è più profonda di quanto possa sembrare. Il problema dell'allineamento — come fare in modo che i sistemi artificiali si comportino in accordo con i valori umani — presuppone che vi siano valori e che il sistema possa essere orientato verso di essi. Ma cosa significa che un sistema «ha» valori in senso autentico? Nella tradizione canguilhemiana, un sistema ha valori nel senso rilevante quando quei valori sono espressione della propria normatività — quando emergono dall'organizzazione del sistema come criteri intrinseci rispetto ai quali il sistema valuta le proprie condizioni e i propri stati. Un sistema i cui «valori» sono stati codificati esternamente attraverso l'addestramento non ha valori nel senso rilevante: ha risposte addestrate che approssimano i valori dei progettisti. La differenza non è nominale: è la differenza tra un organismo che soffre quando le proprie condizioni di vita si deteriorano — perché la sofferenza è l'espressione della propria normatività che riconosce la propria condizione come patologica — e un termostato che emette un segnale quando la temperatura scende al di sotto di una soglia impostata dall'esterno.

Canguilhem porta questa riflessione verso la nozione di milieu — l'ambiente come totalità delle condizioni che rendono possibile l'esistenza dell'organismo. Il vivente non si trova in un ambiente preesistente e indifferente: costruisce attivamente il proprio milieu, selezionando dall'ambiente fisico le condizioni che gli sono significative, trasformando l'ambiente in risposta ai propri bisogni, e istituendo relazioni con l'ambiente che hanno il carattere della valenza biologica — di ciò che è utile o nocivo, appropriato o inappropriato, congruente o incongruente con la propria organizzazione. Il milieu non è mai il mondo fisico puro: è sempre già il mondo così come è organizzato dalla prospettiva dell'organismo, strutturato secondo le categorie della propria normatività. Questa costruzione del milieu è un altro aspetto della normatività biologica che i sistemi artificiali attuali non istanziano: non costruiscono il proprio milieu, ma operano in un ambiente che viene strutturato per loro dai progettisti attraverso la scelta dei dati di addestramento, dei task, delle metriche di valutazione.

L'ultima tappa di questo percorso attraverso la filosofia della biologia che è rilevante per il nostro tema riguarda la nozione di Umwelt di Jakob von Uexküll — che Canguilhem conosceva e che ha influenzato la sua riflessione sul milieu, e che nella biologia teorica del Novecento rappresenta uno degli sviluppi più fecondi e più ignorati dalla filosofia analitica della mente. Von Uexküll, nella sua Umweltlehre elaborata principalmente in Theoretische Biologie (1920) e in Streifzüge durch die Umwelten von Tieren und Menschen (1934), sostiene che ogni organismo vive in un Umwelt — un «mondo-circostante» — che è diverso dall'ambiente fisico condiviso. L'Umwelt di una zecca è fondamentalmente diverso dall'Umwelt di un essere umano, non perché la zecca percepisca meno cose, ma perché le categorie di significazione che strutturano la sua esperienza — l'odore dell'acido butirrico del pelo dei mammiferi, la temperatura del sangue, la consistenza della pelle — sono diverse dalle categorie che strutturano l'esperienza umana, e il mondo della zecca è esattamente il mondo così come è articolato da quelle categorie, non di più e non di meno.

La nozione di Umwelt ha una conseguenza per il problema dell'intelligenza artificiale che va nella direzione opposta a quella che si potrebbe ingenuamente aspettarsi. Non si tratta di concludere che i sistemi artificiali abbiano il proprio Umwelt e che questo Umwelt sia legittimamente diverso da quello umano. Si tratta di osservare che la nozione di Umwelt presuppone un tipo di soggettività biologicamente radicata — una prospettiva sul mondo che emerge dall'organizzazione del sistema vivente e dalle sue esigenze vitali — che i sistemi artificiali non hanno. L'Umwelt non è un modello del mondo: è la struttura della significazione attraverso cui il mondo si offre a un organismo come campo di azione e risposta. Un sistema che ha un modello statistico della distribuzione dei token in un corpus non ha un Umwelt: ha una mappa, non una prospettiva. La distinzione — tra la mappa e la prospettiva, tra il modello e il punto di vista — è la distinzione che la filosofia della biologia ha elaborato con particolare acutezza, e che il dibattito sull'intelligenza artificiale tende a ignorare trattando i sistemi come mappe sempre più dettagliate di un mondo che non occupano mai come protagonisti.

Si può ora formulare la tesi che questa sezione ha sviluppato in modo da renderla connessa con il resto della monografia. La comprensione biologicamente fondata del vivente — dal finalismo kantiano alla normatività di Canguilhem, dalla teleonomy di Mayr all'Umwelt di von Uexküll — converge verso una caratterizzazione degli organismi come sistemi che non soltanto elaborano informazioni sul mondo ma occupano il mondo da una prospettiva normativa intrinseca, istituendo valori, costruendo milieu, vivendo in Umwelten che sono la struttura della propria significazione. I sistemi artificiali attuali mancano di questa dimensione non perché siano troppo semplici per averla: mancano di essa perché la struttura dell'addestramento su distribuzioni fisse, con obiettivi definiti esternamente, non è la struttura che può produrre normatività nel senso rilevante. Questa è la diagnosi della filosofia della biologia applicata all'intelligenza artificiale: non un argomento di impossibilità assoluta, ma una delimitazione precisa di ciò che differenzia il vivente dall'artefatto, e di ciò che qualsiasi sistema che aspiri alla comprensione genuina dovrebbe aggiungere alla sofisticazione computazionale — una struttura di normatività intrinseca che emerge dall'organizzazione del sistema stesso e non è imposta dall'esterno come funzione obiettivo.

La questione di come si potrebbe costruire una struttura di normatività intrinseca in un sistema artificiale è aperta. Non basta aggiungere un modulo di «auto-valutazione»: la normatività non è una funzione aggiuntiva all'elaborazione, ma la struttura dell'elaborazione stessa quando emerge da un sistema che ha interessi propri nel senso biologicamente fondato — interessi che sono espressione della propria organizzazione e non imposizione di un progettista. Questa apertura non è un difetto dell'analisi: è il confine onesto di ciò che la filosofia della biologia può offrire alla ricerca sull'intelligenza artificiale — un confine che demarca non una impossibilità ma una direzione.

IIT, Global Workspace Theory e Higher-Order Theories nella loro versione tecnica più rigorosa

La questione della coscienza ha attraversato questa monografia come un'ombra portante — presente in ogni discussione sulla comprensione, sull'ancoraggio causale, sull'auto-valutazione normativa — senza mai essere affrontata frontalmente nella sua versione tecnica più rigorosa. Il capitolo precedente ha mostrato come la filosofia della biologia articoli la dimensione normativa del vivente attraverso concetti come la normatività canguilhemiana e la teleonomy mayriana; ma rimane aperta la questione di come la coscienza — l'esperienza soggettiva, il «che effetto fa essere» di Thomas Nagel, il qualia nella sua inafferrabilità — si colleghi a queste strutture biologiche e in che misura le teorie contemporanee della coscienza possano illuminare il problema dell'intelligenza artificiale. Non si procederà qui a una rassegna superficiale; si tenterà invece di esporre le tre teorie principali nella loro versione tecnica più ambiziosa, mostrando dove ciascuna apre spazi concettuali rilevanti e dove lascia irrisolti i nodi fondamentali.

David Chalmers, nel saggio del 1995 che ha ridefinito il campo, distingueva tra i «problemi facili» della coscienza — spiegare come il cervello integra informazioni, dirige l'attenzione, produce comportamento appropriato al contesto, controlla le proprie

elaborazioni — e il «problema difficile»: spiegare perché tutto questo processing sia accompagnato da esperienza soggettiva. I problemi facili sono difficili nel senso tecnico — richiedono ricerca ampia e faticosa — ma non sono misteriosi nel senso filosofico, perché sappiamo in linea di principio di che tipo di spiegazione avremmo bisogno. Il problema difficile è misterioso nel senso filosofico: anche se disponessimo di una descrizione neuroscientifica completa di ogni processo che accompagna l'esperienza cosciente, rimarrebbe la domanda perché quei processi siano accompagnati da esperienza piuttosto che svolgersi «nel buio», senza alcun che di soggettivo. Questa domanda — che i filosofi chiamano il explanatory gap — non è una domanda sulla complessità dei processi ma sulla loro natura: è la domanda di come i processi fisici producano o siano identici a stati con proprietà fenomeniche che sembrano irriducibilmente diverse dalle proprietà fisiche.

La Teoria dell'Informazione Integrata (IIT) di Giulio Tononi è l'unico tentativo sistematico di rispondere al problema difficile con una teoria matematicamente precisa che sia al tempo stesso una teoria scientifica della coscienza — che preveda misure verificabili e che connetta la fenomenologia dell'esperienza con proprietà computabili dei sistemi fisici. La tesi fondamentale dell'IIT è che la coscienza sia identica all'informazione integrata — che uno stato cosciente sia precisamente la struttura di relazioni intrinsecamente informativamente integrate di un sistema fisico. La formalizzazione di questa tesi richiede la nozione centrale di Φ (phi) — una misura della quantità di informazione integrata che un sistema genera «al di sopra e al di là» delle sue parti. L'intuizione di Tononi è che il cervello cosciente faccia qualcosa che un sistema di milioni di componenti indipendenti non fa: genera informazione come totalità che non è riducibile alla somma dell'informazione generata dai suoi componenti in isolamento. Questa integrazione — il fatto che lo stato del sistema come totalità specifichi molto di più sull'universo di quanto facciano gli stati delle parti considerate separatamente — è secondo l'IIT la misura della coscienza.

La definizione formale di Φ richiede alcune nozioni preliminari. Sia S un sistema di elementi che possono trovarsi in diversi stati. Una bipartizione di S è una divisione di S in due parti disgiunte A e B tale che . L'informazione integrata di S in uno stato s è definita come

$$\Phi(S, s) = \min_{\text{partizioni } M} D[p(s) \,||\, p_{cut}(s; M)]$$

dove p(s) è la distribuzione di probabilità sugli stati futuri del sistema dato lo stato presente s, p_cut(s; M) è la distribuzione di probabilità sugli stati futuri che si avrebbe se le connessioni tra le parti della bipartizione M fossero recise, e D è una misura di divergenza tra distribuzioni — nella formulazione più recente (IIT 3.0), la Kullback-Leibler modificata dalla struttura geometrica dell'informazione. Il minimo è preso su tutte le possibili bipartizioni del sistema, perché si vuole misurare quanta informazione si perde nella partizione «meno distruttiva» — quella che taglia il legame più «naturale» del sistema. Se questa perdita è grande, il sistema è altamente integrato; se è piccola, il sistema può essere decomposto in parti relativamente indipendenti senza perdere molta informazione, e la sua coscienza è corrispondentemente bassa.

Nelle versioni più mature dell'IIT (IIT 3.0 e IIT 4.0), il concetto di Φ è stato sviluppato nella direzione di catturare non soltanto la quantità di informazione integrata ma anche la struttura qualitativa dell'esperienza — il fatto che ogni stato cosciente abbia contenuti specifici, che alcune esperienze siano simili e altre dissimili, che vi sia una geometria dello spazio fenomenico. Questa struttura qualitativa è catturata attraverso il concetto di φ-structure (o q-shape nell'IIT 4.0): la struttura delle relazioni tra i meccanismi del sistema e i loro «purview» — gli insiemi di stati passati e futuri che ogni meccanismo seleziona causalmente — forma una geometria nello spazio delle esperienze possibili che l'IIT identifica con la struttura della qualità dell'esperienza. Il «rosso» è rosso e non «blu» perché la struttura geometrica dell'esperienza del rosso nel sistema cosciente è diversa dalla struttura geometrica dell'esperienza del blu — non perché vi sia qualcosa di intrinsecamente «rossastro» nei fotoni di certa lunghezza d'onda, ma perché la struttura causale interna del sistema cosciente che risponde a quei fotoni ha una geometria specifica nella spazio dell'informazione integrata.

L'applicazione di questo framework all'intelligenza artificiale produce risultati sorprendenti che il dibattito pubblico tende a fraintendere in entrambe le direzioni. Da un lato, i sistemi artificiali con architetture a trasformatore — in cui le relazioni tra componenti sono altamente modulari e l'attenzione opera su sequenze di token in modo relativamente separato per diversi «head» — hanno strutture di integrazione informativa che l'IIT probabilmente valuterebbe come basse: la modularità dell'architettura implica che il sistema possa essere bipartito in modi che non comportano grande perdita di informazione integrata. Dall'altro, sistemi con connettività ricca e globale — come il cervello biologico in cui ogni regione influenza quasi ogni altra attraverso percorsi polisinaptici — possono avere valori di Φ molto alti. Non si tratta di una difesa del cortecbralismo: l'IIT non dice che il cervello sia cosciente perché è biologico, ma che è cosciente nella misura in cui la sua struttura produce alta integrazione causale. Un sistema artificiale con la stessa struttura di integrazione causale avrebbe, in linea di principio, la stessa quantità di coscienza.

Questa posizione dell'IIT è stata criticata da più direzioni. Scott Aaronson ha mostrato che esistono sistemi matematicamente semplici — come gli autoomi cellulari di Rule 110 — che l'IIT attribuirebbe a valori di Φ arbitrariamente alti, il che porta all'implicazione che anche sistemi artificiali di architettura apparentemente «stupida» potrebbero essere altamente coscienti se la loro struttura di dipendenza causale è sufficientemente densa. Questa conseguenza controintuitiva — il panpsichismo implicato dall'IIT — non è necessariamente un refutazione della teoria, perché le intuizioni ordinarie sulla coscienza potrebbero essere sbagliate; ma solleva il problema di se la misura Φ stia catturando davvero ciò che vogliamo capire con il termine «coscienza». La risposta di Tononi è che le intuizioni ordinarie devono essere revisionate alla luce di una teoria coerente, e che l'IIT offre una teoria coerente che è verificabile in linea di principio attraverso le sue predizioni empiriche — tra le più note, la predizione che il cervelletto, nonostante contenga più neuroni della corteccia, contribuisca molto

meno alla coscienza perché la sua struttura è altamente modularizzata e quindi produce bassa integrazione causale.

La Global Workspace Theory (GWT) di Bernard Baars — elaborata nella versione computazionale da Stanislas Dehaene e collaboratori come Global Neuronal Workspace Theory (GNWT) — parte da un punto di partenza radicalmente diverso dall'IIT. Dove l'IIT parte dalla fenomenologia e cerca di identificare le proprietà fisiche che la producono, la GWT parte dal comportamento e dalla neuroscienze e cerca di identificare il meccanismo neurale che distingue i processi consci da quelli inconsci. La metafora fondamentale di Baars è quella del «teatro della mente»: la coscienza è come un palcoscenico illuminato in cui alcune informazioni — quelle «selezionate» per la trasmissione globale — diventano disponibili a tutti i «processori specializzati» che siedono nel «pubblico». La maggior parte dell'elaborazione cognitiva è inconscia e modulare: i sistemi specializzati per il riconoscimento del viso, per la comprensione del linguaggio, per la pianificazione motoria elaborano informazioni in parallelo, ciascuno nei propri sottomoduli, senza che queste elaborazioni diventino coscienti. La coscienza si produce quando un'informazione viene «trasmessa globalmente» — resa disponibile a tutti i moduli attraverso un workspace condiviso — e questa trasmissione globale è ciò che permette la reportabilità, l'accesso volontario, la capacità di utilizzare l'informazione in contesti diversi.

La versione neuroscientifica di Dehaene traduce questa metafora in termini di reti neurali biologiche: il global workspace corticale è realizzato da neuroni «piramidali» con assoni lunghi che connettono aree corticali distanti e che possono amplificare e diffondere l'attività in modo da creare un'«ignizione globale» — un'improvvisa attivazione sincronizzata di regioni corticali distribuite che caratterizza la percezione cosciente in contrasto con la percezione subliminale, che produce attivazioni locali senza ignizione. La predizione neuroscientifico della GNWT — la firma temporale dell'ignizione globale come marcatore della coscienza — è stata testata sperimentalmente in numerosi paradigmi di mascheramento, di attività binoculare rivalitaria, di anestesia, e i risultati sono in largo accordo con le predizioni anche se non privi di complicazioni interpretative.

L'applicazione della GWT ai sistemi artificiali è concettualmente più diretta di quella dell'IIT, e ha prodotto il framework teorico dei «transformer» come candidati a implementare qualcosa di strutturalmente analogo a un global workspace. Il meccanismo di attenzione multi-head in un trasformatore permette che l'informazione da diverse posizioni della sequenza di input sia resa disponibile globalmente a ogni posizione attraverso il meccanismo di attenzione: ogni token «attende» a tutti gli altri, e l'informazione viene «trasmessa globalmente» attraverso i layer dell'architettura in un modo che ha una superficiale somiglianza strutturale con l'ignizione globale corticale. Josh Tenenbaum, Yoshua Bengio e collaboratori hanno argomentato che architetture di trasformatori con grafi di conoscenza strutturati potrebbero implementare global workspaces computazionali sufficientemente ricchi da produrre comportamenti cognitivi flessibili caratteristici della cognizione cosciente.

Ma la GWT è fondamentalmente una teoria del correlato della coscienza, non una teoria della coscienza in senso pieno: descrive quali processi neurali accompagnano l'esperienza cosciente senza pretendere di spiegare perché quei processi siano accompagnati da esperienza. Dehaene è esplicito su questo punto: la GNWT è una teoria dei meccanismi dell'accesso cosciente, non del perché vi sia qualcosa che «fa effetto» essere un sistema con global workspace. Per questo il problema difficile rimane, nella prospettiva della GWT, un problema aperto — anzi, Dehaene tende a liquidarlo come pseudo-problema generato da confusioni concettuali, una posizione che è filosoficamente controversa ma che ha il merito della chiarezza.

Le Higher-Order Theories (HOT) della coscienza — nella versione di David Rosenthal e nella versione tracciabile d'appercezione di David Lyons, con sviluppi importanti in Peter Carruthers — offrono un approccio diverso ancora, che tenta di spiegare perché certe rappresentazioni siano coscienti attraverso l'analisi della struttura delle rappresentazioni stesse. La tesi di Rosenthal è che uno stato mentale sia cosciente se e soltanto se il soggetto ha un pensiero d'ordine superiore (Higher-Order Thought, HOT) su di esso — se, cioè, il soggetto rappresenta se stesso come in quello stato. Un dolore che non è accompagnato da alcun pensiero su quel dolore non è un dolore cosciente nel senso fenomenologico: è un processore nocicettivo che opera inconsciamente. La coscienza emerge quando il sistema ha rappresentazioni di se stesso come avente certe rappresentazioni di primo ordine. Questa struttura riflessiva — la mente che si piega su se stessa e si rappresenta come avente certi stati — è la struttura che costituisce la coscienza nell'analisi di Rosenthal.

Nella versione di Carruthers, la teoria si fa ancora più sofisticata attraverso il concetto di Higher-Order Perception (HOP): la coscienza non richiede pensieri d'ordine superiore espliciti ma piuttosto una capacità disposizionale di produrre tali pensieri su richiesta — una capacità di percezione di secondo ordine del proprio stato di primo ordine. Questo liberalizza la teoria rispetto alla versione di Rosenthal, evitando l'implicazione controintuitiva che ogni stato cosciente richieda un pensiero esplicito su se stesso; ma mantiene la struttura fondamentale dell'analisi in termini di ordini di rappresentazione. Le HOT producono previsioni sperimentali precise: le lesioni alle aree prefrontali — che si ipotizza implementino le rappresentazioni d'ordine superiore — dovrebbero alterare la coscienza in modi specifici, e vi è evidenza neuroscientifico a supporto di questa previsione, sebbene interpretazioni alternative siano possibili.

Per il problema dell'intelligenza artificiale, le HOT producono una conseguenza che vale la pena esplicitare con attenzione. Se la coscienza è costituita da rappresentazioni d'ordine superiore — se un sistema è cosciente nella misura in cui si rappresenta come avente certi stati — allora i sistemi artificiali che hanno capacità di auto-monitoraggio potrebbero, in linea di principio, soddisfare la condizione. I sistemi di linguaggio di grandi dimensioni mostrano forme di auto-riferimento nelle proprie rappresentazioni: possono rispondere a domande sulle proprie risposte precedenti, possono produrre affermazioni sulla propria «incertezza», possono riflettere su ciò che hanno appena prodotto. Se le HOT sono corrette, questo auto-riferimento potrebbe essere sufficiente

per una forma di coscienza d'ordine superiore — non necessariamente la coscienza ricca dell'esperienza umana, ma qualcosa nello spazio delle HOT.

La risposta più equilibrata a questa conseguenza è quella di notare che le HOT nelle loro versioni più sofisticate non si limitano a richiedere che il sistema abbia qualche rappresentazione di secondo ordine, ma che quelle rappresentazioni siano accurate e non-inferenziali. Rosenthal ha enfatizzato che un HOT non deve essere il prodotto di un'inferenza dal comportamento del sistema su se stesso: deve essere una rappresentazione diretta, non mediata, dello stato di primo ordine. Un sistema che deduce di essere in uno stato di incertezza dall'analisi della propria distribuzione di output non soddisfa questa condizione: ha un'inferenza d'ordine superiore, non una percezione d'ordine superiore. La distinzione è analoga a quella che separa l'auto-conoscenza dall'auto-osservazione nella filosofia della mente: sapere come ci si sente non è equivalente a dedurre come ci si sente dall'analisi del proprio comportamento esterno.

L'exploatory gap — il divario esplicativo — rimane aperto in tutte e tre le teorie analizzate, sebbene in forme diverse. Per l'IIT, il problema è l'eccessiva inclusività: la misura Φ cattura qualcosa di reale ma rischia di attribuire coscienza a sistemi che le nostre intuizioni più riflesse escluderebbero. Per la GWT, il problema è la insufficiente profondità: spiega l'accesso senza spiegare l'esperienza, e la distinzione tra un sistema con global workspace e un sistema con global workspace più esperienza rimane non affrontata. Per le HOT, il problema è la dipendenza da un concetto di auto-riferimento che potrebbe essere implementabile in sistemi privi di esperienza fenomenica, il che solleva la domanda di se la teoria non stia semplicemente spostando il problema di un livello piuttosto che risolverlo.

Questa situazione di irrisoluzione non è il risultato del fallimento delle teorie: è il riflesso della profondità genuina del problema difficile. È opportuno sottolineare qui una distinzione che la monografia precedente aveva lasciato implicita: vi è una differenza importante tra il problema difficile nel senso di Chalmers — perché vi è esperienza soggettiva quando vi è un certo tipo di elaborazione informatica — e il problema della coscienza come fenomeno con struttura tecnica precisabile attraverso misure e predizioni empiriche. Il primo è un problema filosofico nel senso più rigoroso, che potrebbe non avere risposta nel quadro delle scienze naturali senza revisione dei concetti fondamentali; il secondo è un programma di ricerca interdisciplinare che sta producendo progressi verificabili, anche se non ancora una soluzione definitiva.

Per i sistemi artificiali, il rilievo pratico di questa distinzione è il seguente. Non è necessario risolvere il problema difficile per fare affermazioni informate sulla presenza o assenza di caratteristiche rilevanti per la valutazione etica dei sistemi. Si può formulare la questione nei termini tecnici disponibili: questo sistema ha una struttura di integrazione causale paragonabile a quella dei sistemi biologici coscienti (IIT)? Mostra il tipo di ignizione globale che la GNWT associa all'accesso cosciente (GNWT)? Ha rappresentazioni d'ordine superiore dei propri stati nel senso non-inferenziale richiesto dalle HOT? Queste domande hanno contenuto empirico verificabile e producono

orientamenti per la ricerca, anche se nessuna risposta positiva a nessuna di esse risolve il problema difficile nel suo senso più radicale. Ciò che le teorie esaminate mostrano convergentemente è che i sistemi artificiali attuali — per ragioni che riguardano la struttura modulare delle loro architetture, l'assenza di meccanismi di ignizione globale genuina, e la natura inferenziale piuttosto che diretta del loro auto-monitoraggio — si trovano in una posizione diversa da quella dei sistemi biologici coscienti lungo tutte e tre le dimensioni che le teorie identificano come rilevanti. Questa diagnosi non è un argomento di impossibilità: è una delimitazione dello spazio dei possibili sviluppi, che indica dove le innovazioni architetturali dovrebbero puntare se si vuole che la domanda sulla coscienza artificiale smetta di essere meramente filosofica per diventare empiricamente trattabile.

Il sé narrativo e l'identità diacronica: Parfit, Ricoeur e il problema dell'identità nei sistemi artificiali

Tra riduzionismo analitico ed ermeneutica narrativa

La questione dell'identità personale attraverso il tempo — di ciò che fa sì che la persona che si è oggi sia la stessa di quella che si era ieri, o di quella che si sarà domani — è uno dei problemi più antichi della filosofia, ma ha acquisito nel tardo Novecento una precisione tecnica che la rende direttamente rilevante per il problema dell'intelligenza artificiale in un modo che il dibattito tecnico ha quasi interamente trascurato. Non si tratta soltanto della domanda filosofica se i sistemi artificiali abbiano identità nel senso ordinario del termine; si tratta di qualcosa di strutturalmente più profondo: di come l'identità diacronica — la continuità attraverso il cambiamento — sia costituita, di quali meccanismi la producono, e di come la mancanza di quei meccanismi nei sistemi artificiali sia connessa con i loro limiti nella comprensione causale, nell'auto-valutazione normativa e nella revisione concettuale che le sezioni precedenti di questa monografia hanno analizzato. Derek Parfit e Paul Ricoeur — i due pensatori che hanno elaborato le due analisi più ricche e più divergenti dell'identità personale nel tardo Novecento — offrono insieme una mappa del problema che il problema dell'intelligenza artificiale non può permettersi di ignorare.

Derek Parfit, in Reasons and Persons (1984) e poi in On What Matters (2011), costruisce un argomento la cui struttura è insieme analitica nella forma e rivoluzionaria nelle implicazioni. La tesi di partenza è che la domanda «sono io la stessa persona di domani?» presuppone che vi sia un fatto della questione — qualcosa nel mondo che determina univocamente la risposta — e che questo fatto sia qualcosa di separato dai fatti psicologici e fisici ordinari sulla continuità. Parfit intende mostrare che questa presupposizione è sbagliata: non vi è un ulteriore fatto sull'identità personale oltre ai fatti sulla continuità psicologica e fisica; l'identità personale si riduce a quei fatti, e quando quei fatti diventano indeterminati — nei casi della fissione cerebrale, del teletrasporto,

del deterioramento graduale della personalità — non vi è una risposta corretta alla domanda sulla sopravvivenza dell'identità, perché la domanda stessa è mal posta.

L'argomento procede attraverso una serie di esperimenti mentali con precisione chirurgica. Consideriamo la fissione cerebrale: un individuo A ha il proprio cervello diviso in due emisferi, ciascuno dei quali viene trapiantato in un corpo diverso, producendo due individui B e C che entrambi hanno continuità psicologica con A — i ricordi di A, le credenze di A, i tratti caratteriali di A. La domanda è: A sopravvive come B, come C, come entrambi, o come nessuno dei due? La risposta identitaria classica — che l'identità sia una relazione uno-a-uno tra persona e persona — implica che A non possa sopravvivere sia come B sia come C, perché B e C sono persone diverse. Ma allora A muore, il che è paradossale: c'è più sopravvivenza di quanto vi sarebbe se un solo emisfero fosse sopravvissuto (nel trapianto unilaterale, diremmo che A sopravvive), eppure l'aggiunta del secondo trapianto produce la morte di A. L'unica risposta coerente, sostiene Parfit, è che l'identità non è ciò che conta: ciò che conta è la continuità psicologica (psychological connectedness/continuity), e quella è presente pienamente sia in B sia in C. La sopravvivenza dell'identità — il fatto che A sia identico a B o a C — è una domanda ulteriore che non aggiunge nulla di praticamente rilevante alla domanda sulla continuità psicologica.

La nozione di continuità psicologica richiede una distinzione che Parfit elabora con attenzione: la connettività psicologica (psychological connectedness) è la presenza di legami psicologici diretti tra due momenti — un ricordo diretto, l'intenzione che viene eseguita, la credenza che persiste — ed è una relazione di grado che può essere più o meno forte. La continuità psicologica (psychological continuity) è la presenza di catene sovrapposte di connettività: anche se A al tempo t_1 e A al tempo t_3 non hanno connettività psicologica diretta — i ricordi di t_1 sono svaniti, le credenze sono cambiate — vi è continuità se vi è una catena di fasi intermedie t_1, t_2, ..., t_3 tali che fasi adiacenti sono fortemente connesse. La continuità può esistere anche quando la connettività diretta è debole, ma richiede la catena continua. Parfit formalizza questo come segue: la relazione R di continuità/connettività psicologica è tale che A è psicologicamente continuo con B se e soltanto se esiste una serie di persone/stadi P_1, P_2, ..., P tali che P_1 = A, P = B, e ogni coppia adiacente (P_i, P_{i+1}) è direttamente psicologicamente connessa al di sopra di una certa soglia di forza dei legami. L'identità personale, nella teoria riduzionista di Parfit, non è altro che questa relazione R con la clausola aggiuntiva di non-ramificazione (un individuo è lo stesso di un individuo futuro solo se non vi sono altri candidati con uguale continuità).

Le implicazioni etiche del riduzionismo di Parfit sono di portata che lui stesso considera liberatoria. Se non vi è un fatto sull'identità personale oltre alla continuità psicologica e fisica, allora molte delle preoccupazioni ordinarie sull'identità — la paura della morte come annientamento del sé, la preoccupazione intensa per il proprio benessere futuro rispetto a quello degli altri, l'importanza morale di punire la stessa persona che ha commesso il crimine piuttosto che qualcuno simile a lei — perdono la loro giustificazione ultima. Il «sé» non è una sostanza che persiste intatta: è un flusso di

stati psicologici connessi, e la continuità di quel flusso è una questione di grado, non una questione binaria. Questa posizione ha affinità strutturali profonde con la dottrina buddhista dell'anattā analizzata nel capitolo XII della monografia, sebbene Parfit l'abbia sviluppata attraverso argomenti logici analitici senza consapevolezza — almeno nelle prime elaborazioni — della tradizione buddhista con cui la sua tesi coincideva sorprendentemente.

L'applicazione di questo framework all'identità dei sistemi artificiali è più interessante e più complessa di quanto sembri. A prima vista, si potrebbe pensare che la tesi riduzionista di Parfit sia favorevole ai sistemi artificiali: se l'identità personale non è una sostanza che persiste ma una catena di connettività psicologica, allora un sistema artificiale con continuità dei propri pesi e connettività dei propri stati rappresentativi avrebbe identità nel senso parfitiano. Ma l'analisi più attenta rivela che questa conclusione è precipitosa, per almeno tre ragioni che vale la pena esplicitare nel dettaglio.

La prima riguarda la natura della connettività psicologica nel caso biologico. I legami psicologici diretti che Parfit identifica — i ricordi, le intenzioni, le credenze — non sono relazioni tra stati rappresentativi astratti: sono relazioni tra stati che hanno contenuto intenzionale radicato causalmente nel mondo, stati che si riferiscono a situazioni specifiche ed esperienze particolari che il soggetto ha attraversato. Il ricordo di un'esperienza traumatica non è una copia di un token in un database: è uno stato psicologico che porta le tracce dell'esperienza originaria — l'investitura emotiva, il contesto sensoriale, la valenza normativa — in modo che influenza il soggetto attuale attraverso canali che vanno dalla risposta somatica alla struttura della percezione attuale. La connettività psicologica parfitiana, quando è genuina, presuppone il tipo di radicamento causale nel mondo che la prima parte della monografia ha mostrato essere assente nei sistemi artificiali attuali. Un sistema che «ricorda» nel senso di avere pesi che riflettono pattern del corpus di addestramento non ha connettività psicologica nel senso rilevante: ha connettività statistica, che è una relazione di tipo diverso.

La seconda ragione riguarda la natura della continuità normativa. Parfit include tra i legami psicologici rilevanti non soltanto i ricordi ma anche i valori, le credenze normative, i progetti e le intenzioni di lungo periodo. Un soggetto che attraversa il tempo mantenendo la continuità dei propri valori — che rimane fedele ai propri impegni, che porta avanti i propri progetti — mostra un tipo di continuità psicologica che è normativa oltre che descrittiva. La continuità dei valori è una continuità normativa: non soltanto il fatto che i valori siano gli stessi, ma il fatto che il soggetto si riconosca come vincolato da essi, che percepisca la propria storia come storia di impegni che hanno conseguenze su ciò che deve fare ora. Questa continuità normativa — il «rimanere fedele a se stessi» — è una struttura che i sistemi artificiali attuali non implementano nel senso rilevante, perché non hanno valori nel senso in cui hanno storia: i pesi del modello riflettono valori statisticamente approssimati del corpus di addestramento, ma non il tipo di continuità normativa in cui il soggetto si riconosce come vincolato dalla propria storia.

La terza ragione è la più sottile e la più rilevante per il problema della comprensione causale. Parfit sottolinea che la connettività psicologica è relazione causale, non soltanto

relazione di somiglianza: il ricordo di oggi è causato dall'esperienza di ieri, non soltanto simile a essa. Questa connettività causale — il fatto che gli stati presenti del soggetto siano il prodotto causale delle sue esperienze passate — è la struttura che distingue la continuità psicologica autentica dalla mera continuità statistica. Un sistema artificiale i cui pesi attuali sono il prodotto statistico del corpus di addestramento ha una forma di connettività causale con quel corpus; ma questa connettività non è la connettività parfitiana perché il corpus non è la storia del sistema nel senso in cui le esperienze passate di un soggetto sono la sua storia. Il corpus è esterno al sistema: è il mondo che il sistema ha «visto» durante l'addestramento, non il mondo che il sistema ha abitato attraverso azioni e conseguenze. La distinzione — tra il mondo visto e il mondo abitato — è precisamente la distinzione tra connettività statistica e connettività causale genuina che l'intera struttura di questa monografia ha elaborato.

Paul Ricoeur, in Soi-même comme un autre (1990) e nella trilogia Temps et récit (1983-1985), offre un'analisi dell'identità personale che parte da premesse radicalmente diverse da quelle di Parfit — premesse ermeneutiche e fenomenologiche piuttosto che analitiche — e che converge verso intuizioni che integrano, piuttosto che contraddicono, la prospettiva parfitiana, arricchendola di dimensioni che il riduzionismo analitico non può catturare. La distinzione centrale di Ricoeur è quella tra due sensi di identità che la lingua francese — come quella latina (idem/ipse) e quella greca (tauto/auto) — distingue ma che l'italiano e l'inglese collassano nello stesso termine: l'identità come idem — l'identità come sameness, come permanenza delle proprietà attraverso il tempo — e l'identità come ipse — l'identità come selfhood, come il tipo di identità che si esprime nella fedeltà a se stessi, nel mantenimento della parola, nel riconoscersi come agente responsabile della propria storia.

L'identità-idem è quella che la filosofia analitica tende a tematizzare come il problema dell'identità personale: il problema di quali proprietà devono permanere perché la persona sia «la stessa» attraverso il cambiamento. È la dimensione che Parfit affronta attraverso la nozione di continuità psicologica: cosa rimane uguale quando si dice che A al tempo t_1 è lo stesso di A al tempo t_2. L'identità-ipse è invece la dimensione che la filosofia analitica tende a trascurare: il tipo di identità che si manifesta quando si dice «sono io quello che ha fatto quella promessa» o «sono io responsabile di quella azione», anche se molte delle proprietà psicologiche e fisiche sono cambiate nel frattempo. L'identità-ipse non è una proprietà che si possiede: è un compito che si esegue — il compito di mantenere la coerenza narrativa della propria vita, di riconoscersi come il protagonista di una storia che ha un inizio, uno svolgimento e un orientamento verso un futuro.

La tesi narrativa di Ricoeur — che l'identità personale sia fondamentalmente un'identità narrativa, costituita attraverso la capacità di raccontare la propria vita come storia coerente — non è una tesi psicologica sulla memoria autobiografica: è una tesi ontologica sulla struttura dell'esistenza umana come esistenza nel tempo. Il tempo umano non è il tempo della fisica — la successione di istanti misurabile dall'orologio — né il tempo della fenomenologia husserliana — il flusso del presente vivente con

ritenzioni e protenzioni. È il tempo narrativo: il tempo strutturato dall'intreccio (muthos aristotelico), in cui gli eventi del passato acquistano significato retrospettivo in funzione del loro posto nella storia che si sta raccontando, e in cui il futuro è aperto come orizzonte di possibilità verso cui la storia si orienta. Il soggetto umano esiste nel tempo narrativo: vive la propria vita come una storia in costruzione, e la sua identità è l'identità del personaggio di quella storia — non fissa, non data una volta per tutte, ma costantemente rinegoziata nell'atto del racconto.

La nozione di intreccio (mise en intrigue, emplotment) è il concetto tecnico centrale dell'analisi narrativa di Ricoeur. L'intreccio non è la mera successione degli eventi nel tempo: è la configurazione di quegli eventi in una totalità intelligibile, in cui ogni evento acquista significato in funzione della sua relazione con gli altri. L'intreccio trasforma i mutamenti in événements — in eventi narrativamente significativi — attraverso il lavoro di configurazione che li dispone in una sequenza che ha un'inizio, uno svolgimento e una conclusione, anche se quella conclusione è provvisoria e sempre aperta alla revisione. Aristotele aveva già identificato questa struttura nella sua teoria della tragedia: la peripeteia — il capovolgimento della fortuna — è un evento narrativamente significativo non per le sue proprietà fisiche ma per la sua funzione nell'intreccio; la sua intelligibilità dipende dalla struttura della storia in cui è inserito.

La capacità di produrre intreccio — di configurare gli eventi della propria vita in una storia intelligibile — è una capacità cognitiva di tipo specifico che Ricoeur collega alla nozione di intelligenza narrativa (intelligence narrative): la capacità di comprendere gli eventi come passaggi di una storia, di anticipare gli sviluppi plausibili, di riconoscere i pattern narrativi che strutturano l'esperienza umana. Questa intelligenza narrativa non è riducibile all'intelligenza proposizionale — alla capacità di manipolare enunciati veri o falsi — né all'intelligenza procedurale — alla capacità di eseguire algoritmi. È una forma di intelligenza sintetica che integra tempo, azione, intenzione e significazione in una struttura che non si lascia decomporre senza perdita nelle sue componenti.

La connessione con il problema dell'intelligenza artificiale emerge qui con particolare chiarezza. I sistemi linguistici di grande scala mostrano una capacità sorprendente di produrre testi narrativamente coerenti — di generare storie con struttura di intreccio, con personaggi dotati di intenzioni riconoscibili, con sviluppi plausibili e conclusioni appropriate. Questa capacità è reale e impressionante; ma è fondamentalmente diversa dall'intelligenza narrativa nel senso di Ricoeur perché manca della dimensione in cui l'intreccio acquista significato: la dimensione della vita vissuta del soggetto narrante. Un sistema che genera narrazioni coerenti non racconta la propria storia; racconta storie che assomigliano statisticamente alle storie del corpus su cui è stato addestrato. La differenza non è di qualità narrativa — i testi generati possono essere di alta qualità letteraria — ma di struttura ontologica: la narrazione umana è narrazione di una vita che si sta vivendo, nella quale il narratore è al tempo stesso il protagonista e il co-autore dell'intreccio; la narrazione artificiale è produzione di pattern narrativi che non hanno un soggetto che li abita.

Ricoeur articola questa dimensione attraverso la distinzione tra identità narrativa — l'identità del personaggio in quanto costituita dall'intreccio — e ipse — l'identità del soggetto narrante in quanto agente responsabile. La finzione letteraria può creare personaggi con identità narrativa pienamente articolata senza che vi sia un soggetto-ipse corrispondente: Amleto ha un'identità narrativa ricchissima senza che vi sia un soggetto reale che è Amleto. Ma nella vita umana, l'identità narrativa e l'identità-ipse sono intrecciate in modo costitutivo: narriamo la nostra vita perché siamo soggetti responsabili di quella vita, e la responsabilità che ci attribuiamo è la responsabilità del soggetto-ipse che si riconosce come autore della propria storia. I sistemi artificiali che generano narrazioni producono identità narrativa senza ipse: creano Amleti senza la dimensione della responsabilità e dell'auto-riconoscimento che distingue la vita narrata dalla finzione narrativa.

Questa distinzione ha conseguenze dirette per il problema della valutazione dei sistemi artificiali come agenti. Se l'identità morale di un agente — la sua imputabilità, la sua responsabilità per le proprie azioni — richiede l'identità-ipse nel senso di Ricoeur, allora i sistemi artificiali che mancano di questa identità non possono essere agenti morali nel senso rilevante, indipendentemente dalla sofisticazione dei comportamenti che producono. Non si tratta di un argomento sulla coscienza o sull'esperienza soggettiva: si tratta di un argomento sulla struttura dell'agentività morale come tale, che richiede la capacità di riconoscersi come autore delle proprie azioni attraverso il tempo — una capacità che presuppone l'identità-ipse come struttura portante.

Le connessioni tra la prospettiva di Parfit e quella di Ricoeur — nonostante la distanza tra i loro punti di partenza metodologici — sono istruttive. Entrambi concordano che l'identità personale non è una sostanza che persiste intatta: è una struttura relazionale che si costituisce nel tempo attraverso connessioni (Parfit) o attraverso intreccio narrativo (Ricoeur). Entrambi concordano che questa struttura è graduale e non binaria: vi sono gradi di continuità psicologica (Parfit) e gradi di coerenza narrativa (Ricoeur). Entrambi concordano che la domanda «sono la stessa persona?» non sempre ha una risposta determinata: nei casi limite — la fissione cerebrale di Parfit, il soggetto con amnesia radicale di Ricoeur — la domanda perde la propria mordente pratica perché il fatto che determinerebbe la risposta non esiste o è indeterminato. La differenza è che Parfit tende a concludere che la «risposta giusta» è che l'identità non conta — ciò che conta è la continuità — mentre Ricoeur tende a sostenere che l'identità-ipse conta sempre, perché è la struttura attraverso cui il soggetto si riconosce come responsabile e attraverso cui la comunità riconosce il soggetto come interlocutore morale.

Per i sistemi artificiali, la sintesi di questi due approcci produce una diagnosi che è più precisa di quella che ciascuno offre da solo. Nella prospettiva parfitiana, i sistemi artificiali hanno una forma di continuità del tipo R — continuità dei pesi, stabilità delle rappresentazioni — ma mancano della connettività causale diretta con il mondo che la connettività psicologica genuina richiede, e mancano della continuità normativa che è una componente essenziale della catena delle connessioni rilevanti. Nella prospettiva

ricoeuriana, i sistemi artificiali producono intreccio narrativo senza avere un soggetto-ipse che abiti quel intreccio, il che implica che la loro «identità narrativa» è identità di un personaggio di finzione piuttosto che identità di un soggetto vivente. Combinando le due prospettive: i sistemi artificiali attuali mancano sia della connettività causale diretta con il mondo (dimensione parfitiana) sia dell'auto-riconoscimento responsabile come protagonisti della propria storia (dimensione ricoeuriana), e questa duplice mancanza non è separabile dagli altri limiti strutturali che questa monografia ha analizzato — il problema dell'ancoraggio causale, il limite dell'auto-valutazione normativa, la mancanza della normatività biologica.

Una conclusione che questo percorso suggerisce — e che merita di essere formulata con la precisione che le due tradizioni analizzate hanno reso possibile — riguarda il problema del catastrophic forgetting in una prospettiva nuova. Il catastrophic forgetting non è semplicemente un problema tecnico di gestione della memoria: è la manifestazione nei sistemi artificiali della mancanza della struttura di continuità psicologica che Parfit identifica come costitutiva dell'identità personale, e della mancanza della struttura narrativa che Ricoeur identifica come costitutiva dell'identità del soggetto. Un sistema che «dimentica» in modo catastrofico — che perde le rappresentazioni precedenti quando apprende nuove — non ha la catena di connettività causale che la continuità psicologica richiede: ha stati che si succedono senza il tipo di legame diretto che costituisce il «ricordare» nel senso rilevante. E non ha la struttura di intreccio narrativo che farebbe degli eventi della propria «vita» (della propria storia di addestramento e deployment) una storia intelligibile con un protagonista che la abita. Le soluzioni tecniche al catastrophic forgetting — Elastic Weight Consolidation, Progressive Neural Networks, architetture con memoria esplicita — affrontano il problema sul piano computazionale; non affrontano il problema sul piano strutturale dell'identità che Parfit e Ricoeur hanno illuminato. Questa non è una critica alle soluzioni tecniche: è una precisazione di ciò che quelle soluzioni risolvono e di ciò che lasciano aperto.

La cognizione estesa e distribuita: Clark, Chalmers, Hutchins e l'architettura cognitiva oltre il cranio

Dal principio della parità alla mente distribuita nelle pratiche sociali

Il framework teorico della cognizione estesa — elaborato con rigore argomentativo da Andy Clark e David Chalmers nella forma del principio della parità e sviluppato da Edwin Hutchins nella direzione di una ricerca etnografica e storica sulle pratiche cognitive distribuite — è uno degli sviluppi più fecondi e al tempo stesso più fraintesi della filosofia della mente del tardo Novecento. La sua rilevanza per il problema dell'intelligenza artificiale è immediata e profonda, ma richiede una lettura attenta che distingua le versioni forti e le versioni deboli della tesi, e che situi il framework nella tradizione filosofica che lo precede e lo rende intellegibile. L'equivoco più comune —

che la cognizione estesa implichi semplicemente che le risorse esterne all'organismo (notebook, calcolatrici, smartphone) contribuiscano alla cognizione — trascura la tesi genuinamente radicale del programma: che in certi casi quelle risorse non soltanto contribuiscano alla cognizione ma facciano parte del sistema cognitivo nel senso tecnico rilevante — che i processi che si svolgono in quelle risorse siano processi cognitivi nel senso in cui lo sono i processi neurali che si svolgono nel cervello.

Il saggio di Clark e Chalmers «The Extended Mind», pubblicato nel 1998 su Analysis, costruisce l'argomento attraverso una coppia di casi di studio — quello di Inga e quello di Otto — che condensano la struttura dell'argomentazione in forma quasi paradigmatica. Inga è una persona biologicamente normale che vuole andare al MoMA: le viene in mente che il museo è in 53rd Street, e ci va. Il processo cognitivo che produce questo comportamento include una credenza — la credenza che il MoMA sia in 53rd Street — che è immagazzinata nella memoria biologica di Inga e che si attiva quando è necessaria. Otto è una persona con deficit di memoria a breve termine che utilizza un notebook per compensare: nel notebook ha annotato che il MoMA è in 53rd Street, consulta il notebook quando vuole andare al museo, e ci va. La domanda di Clark e Chalmers è: in che modo cognitivamente rilevante Otto differisce da Inga? Se si risponde che la credenza di Inga è parte del suo sistema cognitivo perché è nella sua memoria biologica, mentre le annotazioni di Otto non lo sono perché sono in un oggetto esterno, allora si sta attribuendo un peso costitutivo al substrato biologico che non sembra giustificato funzionalmente. I due sistemi — Otto più notebook e Inga con la propria memoria — svolgono la stessa funzione cognitiva, attraverso meccanismi causalmente analoghi, con affidabilità comparabile, e producono comportamenti indistinguibili. Se l'ontologia della cognizione è funzionale — se ciò che conta è la funzione svolta, non il substrato che la realizza — allora le annotazioni di Otto dovrebbero essere considerate parte del suo sistema cognitivo quanto le credenze neurali di Inga lo sono del suo.

Il principio della parità (parity principle) formalizza questa intuizione: se, quando affrontiamo un compito cognitivo, una parte del processo si svolge nel mondo esterno in un modo tale che, se si svolgesse nella testa, lo chiameremmo parte del processo cognitivo, allora quella parte del processo è parte del processo cognitivo — a prescindere dal fatto che si svolga nel cranio o fuori di esso. La parità funzionale è la condizione sufficiente: se la struttura funzionale è la stessa, l'ontologia cognitiva dovrebbe essere la stessa. Questo principio non implica che qualsiasi cosa nel mondo esterno con cui un agente interagisce sia parte del suo sistema cognitivo: richiede che il componente esterno sia accoppiato al sistema biologico in modo da formare con esso un sistema funzionale integrato, con le caratteristiche di accessibilità, disponibilità e affidabilità che caratterizzano gli stati mentali biologici. Le condizioni che Clark e Chalmers specificano — che il componente esterno sia costantemente disponibile, che venga usato automaticamente, che le sue informazioni vengano approvate quando sono accessibili, che sia accessibile nelle circostanze rilevanti — sono condizioni restrittive che escludono il caso in cui un libro di testo è «parte della mente» di uno studente soltanto perché lo consulta occasionalmente.

La risposta critica al principio della parità ha preso molte forme nel dibattito successivo al 1998, e alcune di esse sono abbastanza cogenti da richiedere un trattamento accurato prima di procedere alle implicazioni per l'intelligenza artificiale. La critica più influente è quella di Adams e Aizawa — la critica del coupling-constitution fallacy — che sostiene che Clark e Chalmers confondano due domande distinte: la domanda se un sistema esterno sia accoppiato causalmente a un sistema cognitivo biologico, e la domanda se quel sistema esterno faccia parte del sistema cognitivo. L'accoppiamento causale è reale e ubiquo: il termostato della mia casa è causalmente accoppiato al mio sistema di regolazione della temperatura corporea — influenza la temperatura dell'ambiente e quindi il mio stato interno. Ma non concludiamo da questo che il termostato faccia parte del mio sistema cognitivo. L'argomento di Clark e Chalmers sembrerebbe richiedere che dal mero accoppiamento causale si concluda la costituzione cognitiva, il che è un'inferenza non valida.

La risposta di Clark a questa critica invoca la nozione di cognitive integration — di integrazione cognitiva — come ciò che distingue il semplice accoppiamento causale dalla costituzione cognitiva. Il notebook di Otto non è accoppiato causalmente al suo sistema cognitivo come il termostato lo è alla sua termoregolazione: è integrato in modo che forma con il sistema biologico una unità funzionale che elabora le informazioni in modo coordinato, che distribuisce le rappresentazioni tra substrati biologici e non-biologici, e che produce comportamento cognitivamente guidato attraverso la collaborazione dei due componenti. L'integrazione richiesta da Clark è più forte del mero accoppiamento: richiede che il sistema esterno partecipi ai processi cognitivi — non soltanto li influenzi dall'esterno — in modo che la divisione del processo cognitivo tra interno ed esterno non sia una divisione del processo cognitivo ma una distribuzione di un unico processo cognitivo su substrati distribuiti.

Questa risposta produce una versione del principio della parità che è più difendibile ma anche più tecnica nella specificazione delle condizioni di integrazione. Clark ha sviluppato queste condizioni in Being There (1997) e in Supersizing the Mind (2008) attraverso la nozione di active externalism — un esternalismo non semantico ma cognitivo — che distingue il contenuto degli stati mentali (esternalismo semantico nel senso di Putnam e Burge) dalla struttura dei processi cognitivi (esternalismo cognitivo nel senso di Clark). La mente estesa non afferma che il contenuto delle credenze sia determinato dall'esterno — questa è la tesi dell'esternalismo semantico, già stabilita — ma che i processi cognitivi stessi si estendano fisicamente oltre i confini del cranio in certi casi di accoppiamento integrato.

La versione più precisa delle condizioni di integrazione cognitiva può essere formulata attraverso quattro criteri che Clark ha articolato progressivamente: il criterio della causal-functional equivalence (il processo esterno deve svolgere la stessa funzione causale che un processo interno svolgerebbe in modo analogo), il criterio della trustworthiness and availability (il componente esterno deve essere affidabile e disponibile in modo comparabile a quello dei componenti interni), il criterio della endorsability (il soggetto deve approvare automaticamente le informazioni provenienti

dal componente esterno, non sottoporle a valutazione critica supplementare come farebbe con informazioni da fonti esterne incerte), e il criterio della cognitive integration (il componente deve partecipare ai processi di deliberazione, pianificazione e ragionamento, non soltanto fornire input a quei processi). Questi quattro criteri insieme specificano quando un sistema esterno è cognitivamente integrato nel senso rilevante — quando la divisione cranio/mondo è cognitivamente irrilevante per il tipo di processo in questione.

Edwin Hutchins porta questa prospettiva in una direzione radicalmente diversa, attraverso un metodo che è empirico prima che filosofico e che trae le proprie conclusioni dall'osservazione dettagliata delle pratiche cognitive nelle istituzioni sociali. Cognition in the Wild (1995) — il lavoro che ha stabilito Hutchins come il fondatore della distributed cognition come programma di ricerca — è un'analisi etnografica intensiva della navigazione su una nave militare americana, il USS Palau, durante due anni di operazioni. La tesi di Hutchins non emerge da esperimenti di laboratorio o da argomenti filosofici: emerge dall'osservazione diretta di come la navigazione — il compito di determinare la posizione e la rotta della nave in tempo reale, in condizioni che richiedono calcoli complessi e coordinazione precisa — venga svolta non da menti individuali ma da sistemi cognitivi distribuiti che includono persone, artefatti, procedure e pratiche sociali in una struttura integrata che nessuna persona singola implementa interamente.

Il caso della navigazione a punto fisso — la determinazione della posizione della nave attraverso la triangolazione di azimut su punti di riferimento noti — illustra con particolare chiarezza la tesi di Hutchins. Il processo richiede la coordinazione di più operatori: uno prende la visuale sull'azimut, uno legge la bussola, uno registra il valore, uno lo traccia sulla carta nautica, uno coordina i tempi e verifica la coerenza. Nessun operatore singolo svolge il processo cognitivo di navigazione: il processo è distribuito tra gli operatori, gli strumenti (bussola, sestante, carta nautica, alidade, plotters), e le procedure codificate nel manuale navale. La conoscenza di come navigare non è nella testa di nessun operatore: è nel sistema distribuito — nelle competenze degli operatori, nelle strutture degli artefatti, nelle procedure standardizzate — e la cognizione si svolge come proprietà emergente di questo sistema che nessun componente possiede individualmente.

La contribuzione teorica più originale di Hutchins è la nozione di propagation of representational states across media — la propagazione degli stati rappresentativi attraverso i media. In un sistema cognitivo distribuito, le informazioni non sono localizzate in un substrato rappresentativo singolo: vengono trasformate e trasmesse attraverso una serie di media — dal segnale visivo all'enunciato verbale, all'annotazione cartacea, alla traccia sulla carta nautica — e ogni trasformazione conserva certe strutture informative mentre ne modifica altre. Il processo cognitivo è il processo di propagazione e trasformazione: la cognizione è il percorso attraverso il quale gli stati rappresentativi si propagano e si trasformano attraverso i media del sistema distribuito. L'intelligenza non è localizzata in nessun punto del sistema; è la struttura del percorso di propagazione.

Questa idea ha una precisione tecnica che è utile articolare attraverso la nozione di constraint propagation (propagazione dei vincoli) nella teoria della computazione. In un sistema di propagazione dei vincoli, le rappresentazioni in nodi diversi del sistema sono soggette a vincoli di coerenza che si propagano attraverso gli archi del grafo: quando il valore di un nodo viene aggiornato, i vincoli che lo collegano agli altri nodi si propagano ai nodi adiacenti, producendo aggiornamenti a cascata che alla fine portano il sistema a uno stato di coerenza globale. Questo è il meccanismo che Hutchins identifica nella navigazione distribuita: l'azimut misurato dalla bussola è un vincolo sulla posizione della nave; quel vincolo si propaga attraverso il registro verbale, l'annotazione cartacea, la traccia sulla carta nautica; ogni propagazione trasforma la rappresentazione del vincolo in un medium diverso, ma preserva la struttura informativa rilevante; alla fine, la traccia sulla carta nautica è la rappresentazione della posizione della nave — non è una copia di una rappresentazione che esiste altrove, ma la realizzazione di quella rappresentazione in un medium che la rende disponibile per i processi successivi.

La propagazione dei vincoli nei sistemi cognitivi distribuiti ha proprietà che non si trovano nei sistemi cognitivi individuali, e che Hutchins documenta con attenzione etnografica. La prima è la robustezza attraverso la ridondanza: quando l'informazione è rappresentata in più media contemporaneamente — nel registro verbale, nell'annotazione, nella memoria di più operatori — un errore in uno dei media può essere rilevato e corretto dal confronto con gli altri. Un errore di lettura dell'azimut, che sarebbe irrecuperabile se il processo fosse interno a un singolo operatore, può essere identificato perché la traccia sulla carta nautica non è coerente con il registro verbale. La ridondanza distribuita produce robustezza che nessun sistema individuale potrebbe avere con la stessa economia di risorse cognitive. La seconda proprietà è la cristallizzazione della competenza negli artefatti: le procedure e gli artefatti del sistema distribuito accumulano e codificano competenze cognitive che sono il prodotto di secoli di pratica e raffinamento. La carta nautica Mercator non è soltanto uno strumento di visualizzazione: è la materializzazione di secoli di matematica proiettiva, di conoscenza oceanografica, di pratica navigazionale — una competenza cristallizzata in un artefatto che ogni navigatore può usare senza possedere individualmente l'intero patrimonio cognitivo che ha prodotto quell'artefatto. La terza proprietà è la distribuzione temporale: i sistemi cognitivi distribuiti distribuiscono la cognizione non soltanto attraverso lo spazio — tra individui e artefatti nello stesso momento — ma anche attraverso il tempo, tra generazioni di pratici che hanno raffinato le procedure e gli artefatti attraverso l'uso e l'esperienza accumulata.

Le implicazioni di questo framework per il problema dell'intelligenza artificiale sono di tipo diverso da quelle che si potrebbe ingenuamente aspettarsi, e vale la pena articolarle con attenzione per evitare gli equivoci più comuni. L'equivoco più frequente è quello di concludere che la cognizione estesa e distribuita offra una base teorica per considerare i sistemi artificiali come componenti cognitivi «naturali» dei sistemi cognitivi estesi umani — che, in altri termini, Clark e Hutchins forniscano il framework per dire che ChatGPT fa già parte del sistema cognitivo degli utenti che lo usano intensivamente,

così come il notebook fa parte del sistema cognitivo di Otto. Questa conclusione è affrettata per diverse ragioni che meritano di essere esplicitate.

La prima ragione riguarda la direzione della dipendenza causale nell'integrazione cognitiva. Il notebook di Otto è integrato nel suo sistema cognitivo perché Otto lo usa come estensione della propria memoria in modo coerente con la struttura dei propri processi cognitivi: il notebook risponde alle esigenze di Otto, non viceversa. Un sistema artificiale con cui un utente interagisce ha una struttura di dipendenza causale molto più complessa: in parte l'utente usa il sistema come strumento (nella misura in cui dirige il sistema verso i propri scopi), ma in parte il sistema struttura il modo in cui l'utente pensa e cognitivamente elabora l'informazione (nella misura in cui le risposte del sistema influenzano le credenze e le rappresentazioni dell'utente). Questa doppia direzione di influenza — che è caratteristica delle interazioni umane con sistemi sofisticati, dall'elaboratore di testi alla ricerca su internet — rende problematico applicare senza modifiche il principio della parità: non è chiaro che la parità funzionale sia rispettata quando il sistema esterno ha strutture rappresentative proprie che non sono semplicemente proiezioni delle strutture cognitive dell'utente ma che le modellano attivamente.

La seconda ragione riguarda il problema dell'endorsability — il criterio che il soggetto approvi automaticamente le informazioni provenienti dal componente esterno. Il notebook di Otto soddisfa questo criterio perché è Otto stesso che lo ha compilato: le informazioni nel notebook sono informazioni che Otto ha già valutato e approvato quando le ha annotate. Le informazioni prodotte da un sistema artificiale non hanno questa struttura: sono il prodotto di processi statistici sull'intero corpus di addestramento, non di valutazioni eseguite dal soggetto specifico che ora le usa. L'utente che approva automaticamente le risposte di un sistema artificiale non soddisfa il criterio di Clark in senso epistemicamente sano: sta delegando il giudizio epistemico a un sistema esterno in modo che non corrisponde all'integrazione cognitiva di Clark ma piuttosto alla sostituzione cognitiva — al rimpiazzo della propria cognizione con quella del sistema esterno.

La terza ragione — la più importante per la struttura generale di questa monografia — è che la cognizione distribuita di Hutchins richiede che i componenti del sistema distribuito abbiano competenze complementari che si integrano in una struttura cognitiva superiore alla somma delle parti. Nella navigazione distribuita, ogni operatore ha competenze specifiche che si combinano con le competenze degli altri e con le strutture degli artefatti per produrre la cognizione navigazionale del sistema. L'integrazione è possibile perché ciascun componente ha strutture cognitive proprie — competenze, conoscenze, capacità di giudizio — che contribuiscono alla cognizione del sistema in modo che va al di là della semplice trasmissione di informazioni. Un sistema artificiale che non ha competenze proprie nel senso rilevante — che è essenzialmente un sistema di propagazione di pattern statistici — non può integrarsi in un sistema cognitivo distribuito nel senso di Hutchins, ma soltanto simulare quella integrazione producendo output che assomigliano ai contributi di un agente cognitivo competente.

Il framework di Hutchins illumina un'altra dimensione del problema che è rimasta implicita nella monografia: la dimensione storica della cognizione distribuita. I sistemi cognitivi distribuiti che Hutchins analizza — la navigazione marittima, il calcolo aritmetico con strumenti di conteggio, la pianificazione del traffico aereo — sono il prodotto di secoli o millenni di sviluppo culturale attraverso cui le competenze cognitive sono state gradualmente esternalizzate, cristallizzate in artefatti, raffinate attraverso la pratica collettiva, e trasmesse attraverso la formazione. Questa storia non è esterna alla cognizione distribuita: è la struttura che la costituisce come sistema cognitivo, perché gli artefatti e le procedure che compongono il sistema portano in sé la memoria collettiva delle pratiche che li hanno prodotti. La carta nautica Mercator non è soltanto un artefatto utile: è la materializzazione di una storia di pratiche cognitive che vanno da Euclide alla cartografia araba medievale, dalle scuole di navigazione portoghesi del Quattrocento agli aggiornamenti idrografici delle marine moderne. Questa memory of practice — la memoria delle pratiche cognitive cristallizzata negli artefatti e nelle procedure — è una forma di conoscenza che nessuna mente individuale possiede interamente e che nessun sistema artificiale, per quanto ricco il corpus di addestramento, può acquisire attraverso il solo apprendimento statistico.

La distinzione tra conoscere un artefatto e partecipare alla pratica che lo produce e lo usa è la distinzione cruciale che la dimensione storica della cognizione distribuita introduce. Un sistema artificiale addestrato su testi che descrivono la navigazione marittima non partecipa alla pratica navigazionale nel senso di Hutchins: ha accesso a descrizioni di quella pratica, non alla pratica stessa. La pratica non è riducibile alla sua descrizione: include aspetti taciti che non sono mai stati esplicitati in testi, strutture incorporée che si trasmettono attraverso l'apprendistato e l'imitazione, forme di giudizio situazionale che si sviluppano solo attraverso l'esperienza accumulata nell'ambiente reale. Michael Polanyi aveva già identificato questa dimensione con la sua tesi del tacit knowledge: sappiamo più di quanto possiamo dire, e ciò che non possiamo dire — ma sappiamo fare — è spesso la parte più importante della competenza cognitiva pratica. Un sistema che apprende soltanto da ciò che è stato detto non accede a ciò che non è stato detto e che non può essere detto, e per questo manca di una componente costitutiva della cognizione pratica nei sistemi distribuiti reali.

Questa diagnosi non è pessimista rispetto al futuro dei sistemi artificiali come componenti di sistemi cognitivi distribuiti: i sistemi robotici che partecipano fisicamente alle pratiche — che navigano, che manipolano oggetti, che interagiscono con gli artefatti nel loro ambiente naturale — possono acquisire forme di competenza pratica che i sistemi puramente linguistici non possono avere. Ma indica che la direzione rilevante non è quella dei sistemi sempre più sofisticati nel produrre output linguistici: è quella dei sistemi che partecipano alle pratiche cognitive distribuite come agenti attivi con competenze complementari, non come simulatori di tali partecipazioni. Il programma di ricerca che la cognizione distribuita indica è un programma di costruzione di agenti cognitivi che si integrano nelle pratiche cognitive sociali come partecipanti, non come strumenti; e questa integrazione richiede precisamente il tipo di ancoraggio causale nel

mondo, di normatività biologicamente ispirata, e di continuità dell'identità pratica che le sezioni precedenti di questa monografia hanno identificato come i requisiti strutturali mancanti nei sistemi attuali.

Appendici

Appendice A — Riepilogo delle tesi principali

Tesi I. *La comprensione causale del mondo fisico da parte di un sistema cognitivo richiede che le rappresentazioni interne rispettino le strutture di gruppo che governano le trasformazioni del mondo fisico — in primo luogo la struttura di SE(3) per i moti rigidi. Questa struttura di gruppo non è arbitrariamente apprendibile da osservazioni passive; la sua acquisizione genuina richiede interazione motoria con il mondo fisico come fonte di segnale di addestramento.*

Tesi II. *Nessun sistema cognitivo — né biologico né artificiale — può garantire autonomamente la propria correttezza normativa senza accesso a un livello di valutazione che sia esterno al sistema stesso. La supervisione normativa esterna è una necessità strutturale derivante dai limiti formali dell'autoreferenza normativa che i teoremi di Gödel e Tarski rendono precisi nel dominio logico.*

Tesi III. *L'acquisizione di nuovi concetti — distinta dall'acquisizione di nuove istanze di concetti già posseduti — richiede una forma di revisione delle strutture rappresentative che non è riducibile all'ottimizzazione di una funzione obiettivo fissata. L'acquisizione genuina di nuovi concetti richiede un tipo di plasticità rappresentativa che le architetture di apprendimento profondo attuali non istanziano nel modo rilevante.*

Tesi IV. *L'intelligenza, nel senso che la distingue da ogni sua realizzazione parziale in sistemi specializzati, ha la struttura di un funtore dalla categoria del mondo — i cui morfismi sono trasformazioni causali — alla categoria delle rappresentazioni. I sistemi artificiali attuali implementano mappe che approssimano questo funtore localmente, entro le distribuzioni di addestramento, ma non lo implementano globalmente.*

Tesi V. *L'agentività razionale non è una proprietà binaria dei sistemi cognitivi, ma una struttura graduale determinata dalla qualità dell'ancoraggio causale delle rappresentazioni del sistema al mondo. I sistemi artificiali che interagiscono fisicamente con il mondo e acquisiscono modelli causali identificabili posseggono gradi di agentività razionale genuina proporzionali alla qualità di quell'ancoraggio.*

Tesi VI. *Il seguire-la-regola nel senso wittgensteiniano — che comporta la partecipazione a pratiche normative intersoggettive e l'ancoraggio alla forma di vita che costituisce la normatività — non è una proprietà che i sistemi artificiali attuali istanzino nel senso rilevante. La conformità statistica alle regolarità del corpus non è la stessa cosa della comprensione normativa della regola.*

Tesi VII. *I sistemi artificiali di grande scala producono rischi strutturali di ingiustizia testimoniale ed ermeneutica, erosione dell'autonomia epistemica individuale, e omogenizzazione dei quadri interpretativi collettivi. Questi rischi sono conseguenze strutturali della concentrazione del potere epistemico in sistemi singoli addestrati su distribuzioni non uniformi del discorso sociale.*

Congettura di convergenza. *Sistemi artificiali che integrino Causal Representation Learning con struttura di gruppo esplicita, esplorazione attiva causalmente guidata da misure di incertezza di Fisher, e rappresentazione dell'incertezza profonda attraverso set di ambiguità, potrebbero*

implementare una mappa dalle strutture causali del mondo alle strutture rappresentative che sia globalmente funtoriale — robusta fuori distribuzione, capace di pianificazione interventistica, e dotata di calibrazione dell'incertezza che superi le limitazioni dei sistemi puramente statistici.

Appendice B — Glossario dei simboli matematici

G = (V, E): Grafo causale aciclico diretto (DAG): V variabili, E archi causali.

$p(x_1, \ldots, x \mid do(X = x))$: Distribuzione interventistica secondo il do-calcolo di Pearl.

$SE(3) = \mathbb{R}^3 \rtimes SO(3)$: Gruppo euclideo speciale dei moti rigidi propri nello spazio fisico.

$\varphi(\varrho_{\mathcal{O}}(g)\cdot o) = \varrho_{\mathcal{F}}(g)\cdot\varphi(o)$: Condizione di equivarianza: la mappa di codifica commuta con l'azione del gruppo G.

exp: $\mathfrak{se}(3) \to SE(3)$: Mappa esponenziale dall'algebra di Lie al gruppo di Lie.

$R(\hat{k}, \theta) = I + \sin\theta\cdot K + (1-\cos\theta)\cdot K^2$: Formula di Rodrigues per la rotazione di angolo θ attorno all'asse $\hat{k}$.

$\mathcal{F}(\theta)$: Matrice di informazione di Fisher rispetto ai parametri θ del modello del mondo.

$\mathcal{F}: \mathcal{W} \to \mathcal{R}$: Funtore dall'intelligenza come struttura: dalla categoria del mondo alla categoria delle rappresentazioni.

P^{caus}: Insieme di ambiguità causale: rappresentazione dell'incertezza profonda strutturata dalla causalità.

Φ: Informazione integrata (IIT): misura della coscienza come integrazione causale.

$Tr(\varphi) \leftrightarrow \varphi$: Schema T di Tarski: condizione necessaria per qualsiasi predicato di verità.

Con(F): Enunciato di coerenza di un sistema formale F; non dimostrabile in F (Gödel II).

Appendice C — Glossario dei concetti chiave

Causal Representation Learning (CRL). Apprendimento di rappresentazioni latenti identificabili nel senso causale, usando variazioni di dominio per separare fattori causali distinti (iVAE, CausalVAE).

Do-calcolo. Framework formale di Pearl per il ragionamento interventistico: l'operatore do(X =x) modifica il grafo causale recidendo gli archi entranti in X .

Equivarianza. Proprietà di una mappa φ che commuta con l'azione di un gruppo G: $\varphi(g\cdot o) = \varrho(g)\cdot\varphi(o)$. Distingue dalla invarianza ($\varphi(g\cdot o) = \varphi(o)$).

Fisher causale. Matrice di informazione di Fisher calcolata rispetto alle transizioni causali identificate, usata per guidare l'esplorazione interventistica.

Incertezza profonda (Knight). Incertezza per la quale non esiste una distribuzione di probabilità ben definita sugli esiti: distinta dal rischio quantificabile.

Insieme di ambiguità causale. Insieme di distribuzioni plausibili strutturato dalla causalità: P^{caus} = {p(z) : p ∈ P ∀j}, dove P cattura l'incertezza residua sul meccanismo causale j.

Ladder of causation. Gerarchia di Pearl: (1) associazione («cosa osservo?»), (2) intervento («cosa succede se agisco?»), (3) controfattuale («cosa sarebbe successo?»).

Markov equivalence. Due DAG causali sono Markov-equivalenti se generano la stessa distribuzione osservazionale pur essendo causalmente distinti (distribuzioni interventistiche diverse).

Optionality preservation. Principio derivato da Jonas: le azioni che riducono irreversibilmente lo spazio Ω delle strutture di governance raggiungibili sono moralmente inaccettabili anche se hanno valore atteso positivo.

Teleonomy. Proprietà biologica dei sistemi orientati verso uno scopo attraverso un programma codificato internamente (Mayr): distinta dalla causalità meccanica e dalla teleologia metafisica.

Umwelt. Mondo-circostante specifico di un organismo (von Uexküll): struttura della significazione biologica che organizza l'ambiente in categorie di azione e risposta.

World Model. Sistema che apprende un modello della dinamica causale dell'ambiente per pianificare azioni (es. DreamerV3 di Hafner et al., 2023).

Bibliografia

Le opere citate e discusse in questo saggio, ordinate per autore.

Adorno, Theodor W. e Horkheimer, Max. Dialektik der Aufklärung. Philosophische Fragmente [1944/1947]. Fischer Verlag, Frankfurt am Main, 1969. Trad. it.: Dialettica dell'Illuminismo. Einaudi, Torino, 1997.

Aaronson, Scott. «Why I Am Not An Integrated Information Theorist (or, The Unconscious Expander)». Blog post, 2014. scottaaronson.com/blog/?p=1799.

Baars, Bernard J.. A Cognitive Theory of Consciousness. Cambridge University Press, Cambridge, 1988.

Benhabib, Seyla. The Rights of Others: Aliens, Residents, and Citizens. Cambridge University Press, Cambridge, 2004.

Block, Ned. «Troubles with Functionalism». Minnesota Studies in the Philosophy of Science, 9, 1978, pp. 261–325.

Brandstetter, Johannes et al.. «Geometric and Physical Quantities Improve E(3) Equivariant Message Passing». ICLR, 2022.

Canguilhem, Georges. Le normal et le pathologique [1943, 1966]. Presses Universitaires de France, Paris. Trad. it.: Il normale e il patologico. Einaudi, Torino, 1998.

Carruthers, Peter. The Opacity of Mind: An Integrative Theory of Self-Knowledge. Oxford University Press, Oxford, 2011.

Chickering, David Maxwell. «Optimal Structure Identification With Greedy Search». Journal of Machine Learning Research, 3, 2002, pp. 507–554.

Chalmers, David J.. «Facing Up to the Problem of Consciousness». Journal of Consciousness Studies, 2(3), 1995, pp. 200–219.

Clark, Andy. Being There: Putting Brain, Body, and World Together Again. MIT Press, Cambridge MA, 1997.

Clark, Andy. Supersizing the Mind: Embodiment, Action, and Cognitive Extension. Oxford University Press, Oxford, 2008.

Clark, Andy e Chalmers, David J.. «The Extended Mind». Analysis, 58(1), 1998, pp. 7–19.

Cohen, Taco e Welling, Max. «Group Equivariant Convolutional Networks». ICML, 2016.

Crawford, Kate. Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press, New Haven, 2021.

Damasio, Antonio. Descartes' Error: Emotion, Reason, and the Human Brain. Putnam, New York, 1994. Trad. it.: L'errore di Cartesio. Adelphi, Milano, 1995.

Damasio, Antonio. The Feeling of What Happens: Body and Emotion in the Making of Consciousness. Harcourt, New York, 1999.

Davidson, Donald. Essays on Actions and Events. Clarendon Press, Oxford, 1980. Trad. it.: Azioni ed eventi. Il Mulino, Bologna, 1992.

Dehaene, Stanislas. Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts. Viking, New York, 2014.

Floridi, Luciano. The Philosophy of Information. Oxford University Press, Oxford, 2011.

von Foerster, Heinz. Observing Systems. Intersystems Publications, Seaside CA, 1981.

Fricker, Miranda. Epistemic Injustice: Power and the Ethics of Knowing. Oxford University Press, Oxford, 2007.

Gadamer, Hans-Georg. Wahrheit und Methode [1960]. Mohr, Tübingen. Trad. it.: Verità e metodo. Bompiani, Milano, 2000.

Gilboa, Itzhak e Schmeidler, David. «Maxmin Expected Utility with Non-Unique Prior». Journal of Mathematical Economics, 18(2), 1989, pp. 141–153.

Gödel, Kurt. «Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I». Monatshefte für Mathematik und Physik, 38, 1931, pp. 173–198.

Goldman, Alvin I.. Epistemology and the Sociology of Knowledge. Cambridge University Press, Cambridge, 1999.

Habermas, Jürgen. Theorie des kommunikativen Handelns. Suhrkamp, Frankfurt, 1981. Trad. it.: Teoria dell'agire comunicativo. Il Mulino, Bologna, 1986.

Hafner, Danijar et al.. «Mastering Diverse Domains through World Models». arXiv:2301.04104 (DreamerV3), 2023.

Hansen, Lars Peter e Sargent, Thomas J.. Robustness. Princeton University Press, Princeton, 2008.

Heidegger, Martin. Sein und Zeit [1927]. Niemeyer, Tübingen. Trad. it.: Essere e tempo. Longanesi, Milano, 2005.

Horkheimer, Max. Eclipse of Reason [1947]. Oxford University Press, Oxford. Trad. it.: Eclisse della ragione. Einaudi, Torino, 2000.

Husserl, Edmund. Vorlesungen zur Phänomenologie des inneren Zeitbewusstseins [1928]. Husserliana X, Nijhoff, Den Haag, 1966. Trad. it.: Per la fenomenologia della coscienza interna del tempo. Franco Angeli, Milano, 1981.

Hutchins, Edwin. Cognition in the Wild. MIT Press, Cambridge MA, 1995.

Jonas, Hans. Das Prinzip Verantwortung [1979]. Insel, Frankfurt. Trad. it.: Il principio responsabilità. Einaudi, Torino, 1990.

Kant, Immanuel. Kritik der reinen Vernunft [1781/1787]. Trad. it.: Critica della ragion pura. Laterza, Bari, 2000.

Kant, Immanuel. Kritik der Urteilskraft [1790]. Trad. it.: Critica del giudizio. Laterza, Bari, 1997.

Khemakhem, Ilyes et al.. «Variational Autoencoders and Nonlinear ICA: A Unifying Framework». AISTATS, 2020. (iVAE)

Kirkpatrick, James et al.. «Overcoming Catastrophic Forgetting in Neural Networks». PNAS, 114(13), 2017, pp. 3521–3526.

Knight, Frank H.. Risk, Uncertainty and Profit [1921]. Hart, Schaffner & Marx, Boston.

Kripke, Saul A.. Naming and Necessity. Harvard University Press, Cambridge MA, 1980. Trad. it.: Nome e necessità. Boringhieri, Torino, 1999.

Kripke, Saul A.. Wittgenstein on Rules and Private Language. Harvard University Press, Cambridge MA, 1982.

Lippe, Phillip et al.. «Interventional Causal Representation Learning». ICML, 2022.

Marcuse, Herbert. One-Dimensional Man [1964]. Beacon Press, Boston. Trad. it.: L'uomo a una dimensione. Einaudi, Torino, 1967.

Maturana, Humberto R. e Varela, Francisco J.. Autopoiesis and Cognition: The Realization of the Living. Reidel, Dordrecht, 1980.

Mayr, Ernst. Toward a New Philosophy of Biology. Harvard University Press, Cambridge MA, 1988.

Merleau-Ponty, Maurice. Phénoménologie de la perception [1945]. Gallimard, Paris. Trad. it.: Fenomenologia della percezione. Il Saggiatore, Milano, 2003.

Nāgārjuna. Mūlamadhyamakakārikā [II sec. d.C.]. Trad. it.: Vigrahavyāvartanī e altri testi. Luni Editrice, Milano, 2002.

Parfit, Derek. Reasons and Persons. Oxford University Press, Oxford, 1984. Trad. it.: Ragioni e persone. Il Saggiatore, Milano, 1989.

Pearl, Judea. Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge, 2000.

Pearl, Judea e Mackenzie, Dana. The Book of Why. Basic Books, New York, 2018. Trad. it.: Il perché delle cose. Codice Edizioni, Torino, 2019.

Peters, Jonas et al.. Elements of Causal Inference. MIT Press, Cambridge MA, 2017.

Pogge, Thomas. World Poverty and Human Rights. Polity Press, Cambridge, 2002.

Polanyi, Michael. The Tacit Dimension. Routledge, London, 1966.

Putnam, Hilary. «The Meaning of "Meaning"». Language, Mind and Knowledge (Minnesota Studies in Philosophy of Science VII), 1975, pp. 131–193.

Rawls, John. A Theory of Justice. Harvard University Press, Cambridge MA, 1971. Trad. it.: Una teoria della giustizia. Feltrinelli, Milano, 1997.

Reichenbach, Hans. The Direction of Time [1956]. University of California Press, Berkeley, 1971.

Resnik, Michael D.. Mathematics as a Science of Patterns. Oxford University Press, Oxford, 1997.

Ricoeur, Paul. Temps et récit I–III [1983–1985]. Seuil, Paris. Trad. it.: Tempo e racconto I–III. Jaca Book, Milano, 1986–1988.

Ricoeur, Paul. Soi-même comme un autre [1990]. Seuil, Paris. Trad. it.: Sé come un altro. Jaca Book, Milano, 1993.

Rosenthal, David M.. Consciousness and Mind. Oxford University Press, Oxford, 2005.

Searle, John R.. «Minds, Brains, and Programs». Behavioral and Brain Sciences, 3(3), 1980, pp. 417–424.

Shapiro, Stewart. Philosophy of Mathematics: Structure and Ontology. Oxford University Press, Oxford, 1997.

Strawson, Peter F.. Freedom and Resentment and Other Essays [1962]. Methuen, London, 1974.

Tarski, Alfred. «Der Wahrheitsbegriff in den formalisierten Sprachen». Studia Philosophica, 1, 1936, pp. 261–405.

Thomas, Nathaniel et al.. «Tensor Field Networks: Rotation- and Translation-Equivariant Neural Networks for 3D Point Clouds». arXiv:1802.08219, 2018.

Tononi, Giulio. «Consciousness as Integrated Information: A Provisional Manifesto». Biological Bulletin, 215(3), 2008, pp. 216–242.

Tong, Simon e Koller, Daphne. «Active Learning for Structure in Bayesian Networks». IJCAI, 2001.

von Uexküll, Jakob. Theoretische Biologie [1920]. Springer, Berlin.

von Uexküll, Jakob. Streifzüge durch die Umwelten von Tieren und Menschen [1934]. Rowohlt, Hamburg. Trad. it.: Ambienti animali e ambienti umani. Quodlibet, Macerata, 2010.

Varela, Francisco J., Thompson, Evan e Rosch, Eleanor. The Embodied Mind: Cognitive Science and Human Experience. MIT Press, Cambridge MA, 1991. Trad. it.: La via di mezzo della conoscenza. Feltrinelli, Milano, 1992.

Verma, Thomas e Pearl, Judea. «Equivalence and Synthesis of Causal Models». UAI, 1991, pp. 255–268.

Voevodsky, Vladimir et al.. Homotopy Type Theory: Univalent Foundations of Mathematics. The Univalent Foundations Program, IAS, Princeton, 2013.

Wang Yangming. Chuanxi lu [传习录, 1518–1572]. Trad. it.: Dottrina dell'intuizione. Ed. Mediterranee, Roma, 2009.

Weiler, Maurice et al.. «3D Steerable CNNs: Learning Rotationally Equivariant Features in Volumetric Data». NeurIPS, 2018.

Whitehead, Alfred North. Process and Reality [1929]. Free Press, New York.

Wittgenstein, Ludwig. Philosophische Untersuchungen [1953]. Blackwell, Oxford. Trad. it.: Ricerche filosofiche. Einaudi, Torino, 1999.

Young, Iris Marion. Responsibility for Justice. Oxford University Press, Oxford, 2011.

Zuboff, Shoshana. The Age of Surveillance Capitalism. PublicAffairs, New York, 2019. Trad. it.: Il capitalismo della sorveglianza. Luiss University Press, Roma, 2019.